AFHRL-TP-89-2

AD-A211 327

AIR FORCE HUMAN RESOURCES LABORATORY

RELATIONSHIPS BETWEEN INDIVIDUAL DIFFERENCES AND ACCURACY IN RATING AIR FORCE JET ENGINE MECHANIC PERFORMANCE

Walter C. Borman
Glenn L. Hallam

Personnel Decisions Research Institute
43 Main Street Southeast, Suite 405
Minneapolis, Minnesota 55414

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

August 1989

Interim Technical Paper for Period December 1986 - March 1989

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFHRL-TP-89-2
6a. NAME OF PERFORMING ORGANIZATION Universal Energy Systems, Incorporated 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Training Systems Division
6c. ADDRESS (City, State, and ZIP Code) 4401 Dayton-Xenia Road Dayton, Ohio 45432-1894	7b. ADDRESS (City, State, and ZIP Code) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b. OFFICE SYMBOL (If applicable) HQ AFHRL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F41689-86-D-0052
8c. ADDRESS (City, State, and ZIP Code) Brooks Air Force Base, Texas 78235-5601	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 62205F; PROJECT NO. 7734; TASK NO 13; WORK UNIT ACCESSION NO. 01

11. TITLE (Include Security Classification)
Relationships Between Individual Differences and Accuracy in Rating Air Force Jet Engine Mechanic Performance

12. PERSONAL AUTHOR(S)
Borman, W.C.; Hallam, G.L.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM Dec 86 TO Mar 89	August 1989	114

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
05	08	
05	09	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
evaluation accuracy; process accuracy; job performance measurement; rater styles; observation accuracy; rating accuracy (Continued)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

It is essential that scorers observe and rate performance accurately when administering work sample tests. In the present research and development effort, videotapes were developed depicting Air Force jet engine mechanics performing work sample tests. The videotapes were used to investigate the accuracy of work sample test scoring, kinds of errors made in the scoring process, the concept of rater styles, and individual difference characteristics hypothesized to predict scorer accuracy. Data were collected from 79 Air Force jet engine mechanics at three Air Force bases. Results indicated first that the observational accuracy index hits (correctly identifying task steps performed properly, i.e., rating "go" performances as "go's") correlated negatively (r = -.40) with the correct rejection accuracy index (correctly identifying task steps performed improperly, i.e., rating "no-go" performances as "no-go's"). Second, more experienced mechanic raters were considerably more critical in their ratings (i.e., made more "no-go" ratings) than their less experienced counterparts. Third, relationships between individual differences variables and observational accuracy were in general low; a perceptual orientation test did, however, correlate significantly (positively) with several of the accuracy indices, and a personality scale measuring

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy J. Allin, Chief, STINFO Branch	22b. TELEPHONE (Include Area Code) (512) 536-3877 22c. OFFICE SYMBOL AFHRL/SCV

DD Form 1473, JUN 86 Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

Item 18 (Concluded):

walk-through performance testing
work sample testing

Item 19 (Concluded):

flexibility correlated positively with hits and negatively with correct rejections. Finally, experience level and some of the individual differences variables correlated significantly with rater style indices such as tendency to nitpick in evaluating others.

AFHRL Technical Paper 89-2 August 1989

RELATIONSHIPS BETWEEN INDIVIDUAL DIFFERENCES AND ACCURACY IN RATING AIR FORCE JET ENGINE MECHANIC PERFORMANCE

Walter C. Borman
Glenn L. Hallam

Personnel Decisions Research Institute
43 Main Street Southeast, Suite 405
Minneapolis, Minnesota 55414

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed and submitted for publication by

Nestor K. Ovalle, 2d, Lt Col, USAF
Chief, Training Assessment Branch

This publication is primarily a working paper. It is published solely to document work performed.

SUMMARY

Research Objectives

There were two major objectives for the present research: (a) to develop videotapes that depict realistic jet engine mechanic performances for future selection and training of hands-on scorers in the Walk-Through Performance Testing (WTPT) program; and (b) to explore individual difference correlates of subject scorer accuracy in rating WTPT performances. The second objective included first assessing the stability or reliability of various observation accuracy component scores, and then examining relationships between these accuracy scores on the part of subject raters and individual differences such as experience as an Air Force mechanic, cognitive ability, mechanical aptitude, and temperament or personality.

Method

The first step in the project was to prepare videotape scripts of six performances on two J-79 installation tasks. Performances were intended to be realistic in that (a) they depicted errors commonly made by first-term Air Force mechanics (e.g., improper safety-wiring procedures); (b) the number of errors scripted into the performances was very close to the number of errors actually made by first-termers performing these tasks in the WTPT program; and (c) the <u>pattern</u> of errors made in the scripts (e.g., the correlation between performances on the two tasks) closely mirrored the pattern of errors in actual WTPT performances. Scripted performances were then videotaped--six mechanics performing each of seven steps on the starter installation task and each of six steps on the collector bowl installation task. Also, a checklist rating form and a special technical order describing proper conduct of each step were prepared for each task

for jet engine mechanic subject raters to use in the rating sessions.

Seventy-nine E-2 to E-7 J-79 mechanics were administered the videotape rating task. They employed the technical orders and checklists to make go or no-go (pass-fail) ratings of each performer on each task step. Also, subjects made summary evaluations of each performer on each task using a 5-point effectiveness scale. The same subjects completed individual differences measures, including temperament scales, background items about experience as a mechanic and as an evaluator, and items tapping rating-related personal characteristics such as self-perceived tendency to nitpick others' performance.

Results

Results showed first that five expert mechanics rating each step of each performance independently and then discussing their checklist ratings to consensus largely confirmed the scripted go and no-go performances. Second, when each subject rater's go/no-go ratings were analyzed against consensus expert judgments according to the observation accuracy components hits (correctly identifying task steps performed properly, i.e., rating "go" performances as go's), and correct rejections (correctly identifying task steps performed improperly, i.e., rating "no-go" performances as no-go's), consistent results were obtained across the two installation tasks. That is, the stability of these observational accuracy scores was reasonably high (reliabilities approximately .60). When evaluation accuracy was computed using the mean expert task-level ratings on the 5-point scales as the target scores against which to compare subjects' evaluation ratings, the resulting differential elevation evaluation accuracy scores were less stable (reliability = .18).

A third noteworthy result was that the two observation accuracy components, hits and correct rejections, correlated negatively ($r = -.40$). Subjects who were relatively accurate on one of these components tended to be relatively inaccurate on the other.

Fourth, relatively experienced raters were considerably more critical in their ratings (i.e., made more no-go ratings) than their less experienced counterparts. This result is in keeping with previous Air Force research (Hedge, Dickinson, & Bierstedt, 1988) and suggests that experience as a mechanic provides a "bias" to err on the side of being critical of mechanic performance because of the extremely high potential human cost of false positives (i.e., rating a "no-go" performance as a "go") compared to false negatives (rating a "go" performance as a "no-go"). High criticalness on the part of more experienced mechanics may be the main reasons for another finding: experience as a mechanic correlated positively with correct rejections ($r = .26$) and negatively with hits ($r = -.42$).

Fifth, relationships between observation accuracy components and evaluation accuracy as indexed in this research were primarily positive but low. Lack of reliability of the evaluation accuracy scores may be the reason for this result.

Sixth, relationships between individual differences variables and observation accuracy were in general low; a perceptual orientation test did, however, correlate significantly (positively) with several of the accuracy indices, and a personality scale measuring flexibility correlated positively with hits and negatively with correct rejections. Finally, experience level and some of the individual difference variables correlated significantly with rater style indices such as tendency to

nitpick in evaluating others.

Recommendations

1. The videotapes developed for the present research should be quite useful for training or possibly selecting WTPT scorers. Regarding training, the tapes can be employed diagnostically before scorer training or sometime during the training process to assess accuracy and rater styles (e.g., not being sufficiently critical). Training can then be focused directly on weaknesses, or, the videotapes might be used as a final examination to evaluate the levels of accuracy achieved at end-of-training. Finally, the tapes can serve as an instructional aid to point out common errors made on the WTPT and how to spot them.

Regarding selection, if there exists a relatively large pool of candidates for WTPT scorer duty, it might be useful to test these candidates on this videotape rating task and select those who receive the highest observation accuracy scores. It is unclear how easily observation skills can be trained; past research suggests such skills may be enhanced to some extent (Thornton & Zorich, 1980). However, training costs may still be reduced by selecting for training those candidates with the best initial observation skills.

2. Overall, individual differences measures in the study showed weak relationships with observation accuracy. It is not advisable at this point to use one or a composite of these individual differences measures to select WTPT scorers. If such selection is to be undertaken, the videotapes themselves should be employed.

PREFACE

The Training Systems Division of the Air Force Human Resources Laboratory is engaged in an effort to develop reliable and accurate performance criteria for use in validating the Air Force selection/classification system, evaluating training programs, and evaluating the quality of other research products. The high-fidelity criterion developed for these purposes utilizes a work sample testing approach called Walk-Through Performance Testing. This paper examines issues related to improving the accuracy of the work sample scoring process. An earlier version of this paper was presented at the annual meeting of the American Psychological Association, Atlanta, 1988.

ACKNOWLEDGMENTS

The authors thank Elaine Pulakos for her considerable help during the planning stages of the research and for her assistance with development of the videotapes. Thanks also to Jerry Hedge, who provided valuable initial guidance in shaping the project and was instrumental in gaining support for the videotape development effort, and to Suzanne Lipscomb for providing wise counsel later in the project and arranging for our Air Force mechanic subjects. Mark Teachout ably guided us through the concluding stages of the project. Finally, we express gratitude to our skilled and knowledgeable technical advisors, MSgts Steve Hotle and Bob Haepanen for their help throughout the project and to Phil Ackerman for his excellent advice on the predictor measures.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLES

TABLE OF CONTENTS (Continued)

FIGURES

APPENDICES

I. INTRODUCTION

Performance evaluation is required for many personnel research applications. For example, accurate measurement of individuals' job performance is critical for evaluating personnel training interventions and for conducting test validation studies to establish effective personnel selection programs.

Previous Air Force Job Performance Measurement Research Relevant to the Present Project

In the Air Force's job performance measurement (JPM) research and development (R & D) program (Hedge & Teachout, 1986), the main objective is to develop a performance assessment system that provides accurate evaluations of individual performance for a wide range of personnel research applications. Within this program, performance is assessed within a number of Air Force specialties (AFSs) by job knowledge tests, performance rating scales, and job task simulations referred to as Walk-Through Performance Tests (WTPTs). The job knowledge tests contain paper-and-pencil, multiple-choice items measuring technical knowledge. The rating scales consist of performance dimensions appropriate for any Air Force enlisted job as well as dimensions developed specifically for individual AFSs. In the WTPTs, examinees perform work samples for tasks; on tasks for which work sampling would be costly or dangerous, examinees are asked to explain what they "would do" to complete steps on the tasks.

As part of the Air Force's JPM program, studies have explored the construct validity of peer, supervisor, and self ratings, and WTPT scores (Kraiger, 1985; Vance, MacCallum, Coovert, & Hedge, 1988). Although results suggest a reasonable degree of convergence among these measures,

there is considerable motivation to enhance their accuracy because of the many important personnel research applications likely to follow from the performance measurement program.

Objectives of the Present Project

One approach to improving the WTPT process is to examine the accuracy with which scorers evaluate work sample performance. A long-term objective of such R & D is to improve the accuracy of work sample scoring. In the present effort, videotapes were developed depicting Air Force mechanics performing work sample tasks. These videotapes were intended to provide realistic stimulus materials for the present research. Another objective of the videotape development work was to provide support for future WTPT scorer training. The videotapes may be used to test for observation skill acquisition on the part of WTPT scorers.

In sum, an investigation was conducted on the accuracy of work sample performance test scoring, along with the kinds of errors made. Individual difference measures hypothesized to predict scorer accuracy were also identified. As described in greater detail below, Air Force enlisted personnel with varying levels of experience as jet engine mechanics viewed the videotapes and made "go" or "no-go" (pass-fail) ratings of performance for each task step. In addition, these participants completed a number of individual difference tests and inventories (biographical, cognitive, and personality measures). Various components of rating accuracy were then computed, and their relationships with the individual differences explored.

Several important scientific questions were addressed as part of this research. These include: (a) How are observation and evaluation

accuracy in performance assessment related? (e.g., Murphy, Garcia, Kerkar, Martin, & Balzer, 1982); (b) What is the utility of employing signal detection theory for the study of observation accuracy? (Baker & Schuck, 1975; Lord, 1985); and (c) What is the relationship between rating _outcome_ accuracy (such as correctly identifying that an error in performance has occurred) and _process_ accuracy (being able to provide the correct reason for each performance error)? These issues are discussed below.

Issues in Assessing Evaluation and Observation Accuracy

In performance rating research, evaluation accuracy is typically operationalized as the degree to which ratings of target ratees agree with some standard or criterion performance levels established by expert judgment (Borman, 1978; Murphy et al., 1982). Observation accuracy has been operationalized as the number of objectively verifiable ratee behaviors correctly noted by the rater (Thornton & Zorich, 1980).

Several studies have investigated evaluation accuracy. For example, Borman (1977, 1978) developed videotapes of actors performing in two job scenarios--one a supervisor in a problem-solving session with a subordinate and the other a recruiting interviewer discussing a job opportunity with a candidate for the job. The videotaped performances were carefully scripted and developed with preset effectiveness levels judged to be realistic by a panel knowledgeable about the jobs.

After the videotapes were developed, 14 industrial psychologists and advanced graduate students viewed the tapes a number of times, studied transcripts of the performances, and then rated the effectiveness of each ratee along several dimensions. Analyses of these expert ratings indicated high agreement among experts, with intraclass correlations on

individual performance dimensions ranging from .91 to .98, with a median intraclass correlation of .97. Further, convergent and discriminant validity (Kavanagh, MacKinney, & Wolins, 1971) indices were high, and agreement between the mean expert ratings and intended target scores was reasonably high (median r for individual dimensions was .91).

Mean expert ratings were used as criterion performance levels against which to assess the evaluation accuracy of subjects viewing the videotapes. In a series of studies, these tapes and mean expert ratings were employed (a) to investigate relationships between performance rating accuracy and various rater errors, such as halo and leniency (Borman, 1977); (b) to explore ability and personality individual difference correlates of raters' accuracy in making performance ratings (Borman, 1979a); and (c) to study the effects of rating format and rater training on rating accuracy (Borman, 1979b).

As mentioned, the above studies dealt with evaluation: viewing a sample of ratee performance, making evaluative judgments about the ratee behaviors exhibited, and integrating those judgments to arrive at a single rating of the ratee on one or more evaluative scales. Other studies have investigated observation: perceiving or failing to perceive a relatively short, discrete sample of ratee behavior. Such studies have also explored the relationship between observation accuracy and evaluation accuracy. In one sense, observation can be viewed as the first step in evaluation.

An example of such an investigation is a study conducted by Murphy et al. (1982) in which students watched a videotape of college instructors delivering lectures. Observation accuracy was operationalized as "the accuracy with which students estimated the frequency of certain performance-related behaviors." Evaluation accuracy

was measured by comparing the overall performance ratings by each student to pooled expert ratings. The two types of accuracy were found to correlate .40 and above for most of the evaluation accuracy indices (elevation, differential elevation, and differential accuracy). This suggests that for those aspects of accuracy, the ability to observe behavior accurately is related to (and possibly a prerequisite for) making accurate integrative summary evaluations of work performance.

Focus of the Present Research: Observation Accuracy

In the present research, raters were required to view a relatively short videotape segment of a ratee performing a task, and then note if the behavior exhibited by the ratee conformed to technical order specifications. In that only go/no-go checklist responses were required, this rating task was more like an observation task than an evaluation task. Accordingly, the present study investigated observation accuracy and the importance of individual differences in making accurate ratings.

In exploring observation accuracy, signal detection theory (Baker & Schuck, 1975; Green & Swets, 1966) provided the basis for defining certain observational errors. Signal detection theory identifies and defines the following components of observation accuracy: Hits (number of correct behaviors properly identified as correct), False Alarms (number of correct behaviors improperly identified as incorrect), Correct Rejections (number of incorrect behaviors properly identified as incorrect, and Misses (number of incorrect behaviors improperly identified as correct). These different observation accuracy components represent quite different observational requirements. Consider, for example, the two observation errors from signal detection theory. False Alarms reflect "reading in" problems, errors, and mistakes where there

are none; Misses, on the other hand, reflect overlooking these kinds of problems and errors when they are present. Therefore, since these elements of observation accuracy may require differing skills and abilities, it is important to examine them individually, using the signal detection theory framework.

The final research issue examined focused on both what can be termed "outcome accuracy" (for example, any of the four components from signal detection theory), and what might be called "process accuracy." Lord (1985) argued that a study of accuracy in evaluating performance can be illuminated by considering both the signal <u>and</u> the noise in a signal detection framework. Signal and noise can be operationalized as Hits and False Alarms, respectively; increasing the signal <u>or</u> decreasing noise will improve accuracy. Attending to both Hits and False Alarms should reveal more details about the rating process with respect to accuracy than if only a summary accuracy measure were used.

Another way to study process accuracy (in contrast to outcome accuracy) is to require a rater, for no-go performances, to provide the proper reason for no-go's. This more stringent accuracy measure is in the spirit of Lord's objective to have accuracy indices that yield greater detail about the rating process. However, this process accuracy measure goes beyond Lord's notion of using Hits and False Alarms for studying rating processes since it requires the rater to know <u>what kind</u> of error or mistake has been made for no-go task step performances.

The present research afforded an opportunity to explore the process accuracy notion by asking raters viewing the videotape to record reasons for their go/no-go ratings. This enabled: (a) an assessment of the empirical relationship between outcome and process accuracy; and (b) an

examination of the possible differences in the patterns of individual difference correlates of process accuracy compared to the correlates of outcome accuracy on the four accuracy components.

In sum, the present R & D effort was intended to develop videotapes of mechanics performing maintenance tasks in order to investigate several specific research questions involving observation accuracy and its individual difference correlates, and to provide a set of stimulus materials that can be used in future research and applications.

II. METHOD

This section first describes the development of rating session materials. These include videotapes depicting Air Force mechanics performing two installation tasks on a J-79 jet engine and the technical orders and rating scales for each of the two tasks. The videotaped performances were scripted such that each of the actors in the videotapes was instructed to make certain mistakes on the installation tasks. The errors selected for scripting into the performances were examples of the kinds of errors typically made on these tasks. Six different videotapes were made of each task; each version depicted the ratee making different errors.

This section next describes procedures that required experienced expert mechanics to study the videotaped performances and arrive at consensus judgments about each actor's performance on each of the steps associated with the tasks. These judges provided separate go/no-go evaluations on each task step for the six videotaped performances within each task, along with an overall evaluation of each actor's performance on each of the two tasks. These go/no-go ratings were expected to confirm the scripted performance levels, converging upon "target scores."

Go/no-go judgments of subjects who later rated actors' performances in the videotapes could then be compared to the target scores in order to assess rating accuracy.

The third part of this section discusses identification and development of individual difference predictor measures hypothesized to correlate with observation or evaluation accuracy. Background experience, cognitive ability, and personality measures identified from past studies or developed specifically for the present effort, comprised a predictor battery of tests and inventories.

Fourth, administration of the predictor battery and videotape performance rating task to 79 Air Force jet engine mechanics of varying experience levels is described. And finally, this section presents details of analyses conducted on these subjects' performance rating data, as well as a description of analysis work done to examine relationships between the individual difference measures and components of performance rating accuracy.

Development of Rating Session Materials

Development of Videotapes

The first step toward developing videotapes of Air Force mechanics working on a jet engine was to select two suitable job tasks. Targeting two tasks for the research allowed assessment of the generalizability of results involving rating accuracy across tasks. Several criteria were employed in selecting the tasks: (a) based on performance data obtained for first-term Air Force mechanics in previous research, there should be variation in performance on the tasks (i.e., some people make mistakes when doing them); (b) the tasks should be suited for depiction by a relatively short videotape; (c) rating performances on the tasks should

be relatively difficult (i.e., raters should vary in the accuracy with which they rate these performances); and (d) the tasks should be suited for division into discrete performance steps. Two tasks that met these criteria were installing a starter and installing a collector bowl on the J-79 jet engine. The J-79 engine was selected because it is a common engine type and many Air Force mechanics have worked on it. Thus, the population of potential rater subjects was relatively large.

Next, the number of task steps and number of ratees were determined. The number of task steps was, in part, dictated by the nature of each task. Extensive consultation with two senior (E-7) jet engine mechanics serving on temporary duty as job analysts resulted in identification of seven steps for each of the two tasks. Brief descriptions of the steps for these tasks are presented in Appendix A. The seventh step of the collector bowl task was subsequently dropped to reduce the length of the videotapes.

It was important to select a sufficient number of videotaped performances (ratees) such that rater accuracy scores obtained from ratings of these performances would be reliable. In previous videotape research, eight actors performed one job and eight other actors were taped performing the other job. The across-job reliability was .46 for subjects' differential accuracy (DA) and .72 for their scores on a halo index (Borman, 1977). These levels appear acceptable but would, of course, decrease when the accuracy scores are computed on smaller numbers of videotaped ratees. In a study using only four of the eight performances for each job, DA across-job reliability was only .20 (Borman & Cascio, 1982). Therefore, taking into account this reliability information and the practical restrictions on length of the performances for subsequent subject viewing, the current study included six actors

performing each of the two tasks.

The major objective in preparing scripts of the videotape performances was to make the performances as realistic as possible. A valuable resource for accomplishing this objective was data that had been collected during another phase of the JPM project on 84 first-term Air Force jet engine mechanics. These mechanics previously had been administered performance tests on the two target tasks (starter and collector bowl) as part of the WTPT for J-79 jet mechanics. Each mechanic had been rated go/no-go on each of the seven steps in these tasks (as well as on several other steps). These data provided base rate information about how often no-go errors are actually made on these two tasks. This information was then used to help guide the total number of go and no-go performances that were scripted into the task steps for the six actors.

In addition, correlations across the steps within the tasks and between the two tasks were computed for the 84 subjects tested previously. These correlations served as targets for scripting go and no-go performances and in developing realistic scripts for the videotapes.

The scripts indicated whether a performance step should be performed correctly or incorrectly and specified the particular performance error to be committed on each step designated as no-go. For example, on one step of the starter task, the actor was instructed to put grease on some, but not all, of the spline teeth of the starter shaft, although the technical order clearly states that the mechanic should put grease on all the splines. The same two experienced E-7 jet engine mechanic advisors referred to above identified the most common errors first-term mechanics make on each step of these tasks, based on the advisors' considerable experience in supervising and observing mechanics performing the tasks.

These intentional no-go mistakes were then incorporated into the scripts. Descriptions of each task step and a summary script of the intended go and no-go performances appear in Appendix A.

For the scripted performances, base-rate and correlational analyses were conducted to assess the match between the intended target score matrix results and results from the 84 first-term mechanics for whom WTPT data were available. Table 1 shows that a reasonably good match was achieved. This, in turn, suggests that when persons familiar with first-term mechanic performance on these tasks view the videotapes, they are likely to concur that the configurations of go and no-go performances are realistic.

At this point, the scripted performances were ready for videotaping. Three Air Force personnel were recruited to appear in the videotapes. Two were first-term (E-3) mechanics and the other was a more experienced (E-7) mechanic. The actual identity of the actors was not important because the camera focused on their hands and arms, not their heads or faces. One of the E-3s acted in three roles for each task (starter and collector bowl). The other E-3 acted in two roles for each task, whereas the E-7 performed as the sixth mechanic in each videotape. All actors were white males.

The videotaping was conducted on a training J-79 engine, which had characteristics and a configuration identical to those typically seen in an operational J-79 repair shop setting. During the 2-day taping session, every attempt was made to adhere to the scripts and to ensure that the taped performances provided a realistic depiction of actual shop

Table 1. Target Score Array Compared to Data from Testing of 84 First-Term Incumbents on These Two Tasks

	Starter task		Collector bowl task	
Data source	Base rate	Median correlation between steps	Base rate	Median correlation between steps
Videotape Intended Target Scores	67% go	.00	56% go	.33
Actual Data	70% go	.11	60% go	.31

Correlation Between Performance on the Two Tasks

Videotape Target Scores:	.41
Actual Data:	.33

performances on the part of first-term mechanics. The videotapes were subsequently edited to eliminate taping flaws and to shorten lengthy safety-wiring sequences.

However, a major objective continued to be to reflect as faithfully as possible for the viewer of the videotapes the experience of watching first-term mechanics complete two tasks. If a particular sequence on the tapes was somewhat long and tedious to watch, the sequence was not edited to shorten it very much, since observing this kind of performance sequence is, in fact, a tedious process. During final editing, graphics displaying the proper step numbers were inserted on the videotapes just before each step was shown to help with subsequent administration of the videotape rating task.

Development of Supporting Materials

Concurrent with developing the videotapes, rating scales and technical orders were prepared for subsequent use in the videotape rating sessions. Rating scales for the starter and collector bowl installation tasks consisted of go/no-go checklist items for each task step for each of the six videotaped ratee performances. A 5-point behaviorally anchored overall performance rating scale for each ratee on each task was also included on the form. All the rating scales appear in Appendix B.

A technical order for each of the two installation tasks was also developed. With the help of the E-7 subject-matter experts, descriptions of how to do each task step were prepared in a format similar to the standard technical orders Air Force mechanics use to compl^te repair and installation tasks. These technical orders were designed to provide basic information for completing the task without specifying every detail to which subject raters should attend. The example in Figure 1 shows two

steps of the starter installation technical order. The numbers in parentheses refer to engine components depicted in an accompanying schematic rendition of the appropriate section of the J-79 engine. The complete technical orders appear in Appendix B along with the rating scales.

Step 2. Place coupling (1) on adapter (2) and properly lock coupling latch to hold coupling on adapter.

Step 3. Raise starter into position and engage starter output shaft (3) with transfer gearbox splines. Rotate starter until breech chamber is at 8 o'clock position and position starter forward, working starter flange (4) under locking edge of coupling.

Figure 1. Two Example Steps from Starter Installation Technical Order.

A pilot test of the videotape rating task revealed certain problems with subjects' interpretations of what was required. For example, some subjects assumed that if they could not actually see part of a step being completed properly, the step should be marked a no-go. This problem surfaced on the collector bowl task when one of the nuts to be tightened was out of camera range and on the reverse side of the engine. This problem, along with a few others, led to development of an "assumption list" as part of the rating task instructions. The list clarified how subjects were to interpret the rating task, and thus helped to standardize the task. The assumption list is also contained in Appendix B.

Expert Judgment Rating Session

After the videotapes were developed and edited, an expert rating session was conducted with five very experienced jet engine mechanics. The expert panel consisted of four E-7s and an E-8, each with at least 10 years of experience in J-79 jet engine maintenance. During the session,

panel members were first briefed on the project and on the rating task, and they carefully reviewed the rating scales and technical orders. Then, they viewed the first ratee performing the first task and each independently made go or no-go ratings of that performance. At that point, panel members discussed the performance as necessary to arrive at a consensus judgment on the step-level ratings. On a number of occasions, the videotape segment was re-run for the panel to help them agree on a rating. The same procedure was followed for each task step performance. In addition, each panel member independently rated the overall task performance of each ratee on each of the two tasks, using the 5-point evaluation rating scale. The final consensus go/no-go ratings and reasons for no-go's appear in Appendix C.

The Predictor and Process Variable Set

A broad array of predictor measures was selected or developed to assess biographical/background information, cognitive abilities, and personality characteristics that past research and current speculation suggest are likely to covary with rater observation accuracy. In addition, several process variable measures were developed and administered as well. This was done because past research indicates the predictors may gain validity through their relationship with process variables, such as knowledge of how the starter and collector bowl should be installed, and motivation to do well in the rating session. Figure 2 provides a list of the predictor and process measures, which were either administered to the study sample of 79 Air Force mechanics or acquired from their personnel files. These measures are described below in greater detail and appear in part in Appendix D.

Biographical Information

Age, rank, experience in Air Force, experience as mechanic, experience as mechanic working on jet engines, experience as mechanic working on J-79 engine, time since last rated another person's job performance, self-rated mechanical ability, educational attainment, and high school grade point average. Also, number of times person has installed a starter on a J-79 engine, installed a collector bowl on a J-79 engine, rated another mechanic's job performance, or rated a non-mechanic's performance.

Knowledges and Abilities

ASVAB Subscales: General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Numerical Operations, Coding Speed, Auto and Shop Information, Math Knowledge, Mechanical Comprehension, Electronics Information

Project A Tests: Orientation Test (24 items), Maze Test (24 items)

Temperament and Rater Style

Self-ratings of detail orientation, tendency to attend to and evaluate the work of coworkers, adequacy of previous night's sleep, tendency to exert effort on tasks, accuracy of first impressions of other people, and tendency to criticize others.

Adjective Check List: Achievement, Dominance, Self-Confidence, Self-Control, Exhibition, Succorance, Abasement, Nurturance, Affiliation, Deference, Autonomy, Aggression, Change, Order, Personal Adjustment, Impulsiveness, and Extroversion.

Post-Rating (Process) Measures: Self-ratings of endurance and motivation for the rating task, knowledge of how starter and collector bowl repair tasks are supposed to be done, the degree to which the individual "nitpicked," adherence to the appropriate technical order, and opportunity afforded to see the videotaped performances clearly.

Figure 2. List of Predictor and Process Variables.

Biographical and Preferences Inventory

This inventory contains 38 items measuring experience as a mechanic; experience in making performance ratings of others; self-perceived mechanical ability; detail orientation; educational achievement (high school grade point average [GPA] and highest level of formal educational attainment); self-perceived effort expended on tasks; attention span; and demographic information such as race, sex, and age.

Tests of Spatial/Perceptual Ability

Two timed tests of spatial/perceptual ability were administered. The first test, Mazes, is a speeded test requiring the examinee to visually scan a maze to identify an entrance to the maze that leads to one of the designated exits. Examinees are given 5 1/2 minutes to complete 24 items. The second test, Orientation, is a power test that requires the examinee to identify what an object will look like when it has been turned or rotated. Examinees are given 10 minutes to complete 24 items. These tests were developed as part of Army Project A, a large-scale effort to evaluate and improve the Army's selection and classification system (Peterson, 1987). In a sample of approximately 9,000 subjects, both tests proved to be highly reliable and to correlate minimally with the Armed Services Vocational Aptitude Battery.

Armed Services Vocational Aptitude Battery (ASVAB)

The ASVAB is a Department of Defense cognitive test battery used to select and classify applicants for military service. The subtests are designed to measure General Science, Word Knowledge, Electronics Information, Mechanical Comprehension, Paragraph Comprehension, Auto/Shop Knowledge, Arithmetic Reasoning, Mathematics Knowledge, etc. The present research focused on two measures: the Mechanical Comprehension subtest

and the Armed Forces Qualification Test (AFQT), a composite of four subtests. The AFQT composite has proven to provide a reasonably good index of general cognitive ability.

Adjective Check List (ACL)

This 255-word check list is an edited version of Gough and Heilbrun's 300-item Adjective List (Gough & Heilbrun, 1965). Forty-five adjectives were removed because they appeared inappropriate for the present study. Temperament scales included in the ACL were: Achievement, Dominance, Self-Confidence, Self-Control, Exhibition, Succorance, Abasement, Nurturance, Affiliation, Deference, Autonomy, Aggression, Change, Order, Personal Adjustment, Impulsiveness, and Extroversion.

Post-Rating Questionnaire

Administered after the rating session, this 27-item questionnaire measures several process variables hypothesized to mediate the relationship between the predictors and rater accuracy. These variables included self-perceptions of initial motivation and endurance on the rating task, knowledge of the correct procedures for completing the starter and collector bowl tasks, and self-perceived ability to accurately perceive ratee behaviors on these tasks. In addition, the questionnaire measures several rater style variables, including the tendency to "nitpick" and the tendency to rely upon the technical order provided as opposed to one's own knowledge of the best methods for completing the maintenance tasks. This instrument was developed specially for the present research effort.

Data Collection and Analysis

Sample

Subjects were 79 Air Force jet engine mechanics ages 19 to 39, ranks E-2 to E-7. They were all presently working on J-79 engines at a shop facility on one of three bases visited by the research staff for data collection. Subjects were assigned to the rating task by the senior NCOs in charge according to instructions to provide mechanics with differing levels of experience as Air Force jet mechanics.

The mean age of the subject raters was approximately 24. All but one rater was male; 76% were White and 14% Black. On the average, respondents had worked as mechanics (civilian and military) for about 6 years and had installed a starter or collector bowl on a J-79 engine about three times. Many individuals, however, had never installed a starter (34%) or a collector bowl (49%). More than 50% had never formally rated the job performance of another mechanic; about 82% had never formally rated the job performance of a non-mechanic. A complete description of the sample appears in Table 2.

Data Collection Procedures

Each data collection session was attended by 7 to 12 Air Force mechanics and lasted approximately 5 hours. Three sessions were conducted at the Air Force base in the South, three at the base in the East, and three at the western base. After receiving an introduction to the study by the experimenters, the participants completed the background information questionnaire, were administered the two timed spatial/perceptual tests, completed the Adjective Check List, and then took a short break.

Table 2. Description of Sample

Age (Mean and SD)	24.5	4.8
Race (Percent)		
Black	13.9	
White	75.9	
Hispanic	7.6	
Asian	1.3	
American Indian	1.3	
Rank (Percent)		
E-2	8.9	
E-3	36.7	
E-4	35.4	
E-5	10.1	
E-6	6.3	
E-7	2.5	
Years in Air Force (Mean and SD)	4.8	4.4
Years as Mechanic (Mean and SD)	6.0	5.4
Years as Air Force Jet Mechanic (Mean and SD)	4.5	4.4
Years Experience with J-79 Engine (Mean and SD)	2.7	2.2

Upon returning from the break, the subjects read the performance rating instructions, were briefed on the rating task, and reviewed the technical orders for the two installation tasks. They then viewed the six videotaped performances. The videotapes were presented such that raters viewed the tapes in a different order in different sessions and neither the two best nor the two worst performances were shown back to back. These procedures were instituted to minimize order and contrast effects. The experimenter stopped the videotape after each step of the performance to provide participants an opportunity to rate the step as go or no-go and to write a brief reason for the rating. For example, when the mechanic on the videotape moved the torque wrench in a "jerky" fashion, many respondents rated the step as a no-go, then wrote "improper use of torque wrench." After rating the final step of each performance, participants rated the overall effectiveness of the task performance, using the 5-point scale appearing in Figure 3. Participants were asked to work independently and were given a break for lunch after viewing the first three ratees. When all videotapes had been viewed and rated, participants completed the post-rating questionnaire and were thanked for their participation.

Low				High
1	2	3	4	5
Mechanic performed the task incorrectly, using improper tools, materials or methods, and completing the job inefficiently or with poor technique				Mechanic completed the job correctly and efficiently, using the appropriate tools and methods, and showing good technique.

Figure 3. Rating Scale for Evaluation of Task Performance.

Data Analysis Procedures

Analysis of Expert Ratings. Expert judge interrater reliability was computed employing an intraclass correlation coefficient for each task step. This measure reflects the agreement across the five experts in their initial pre-consensus rating of the six videotaped ratees on each step (Shrout & Fleiss, 1979). The median intraclass correlation across the 13 ratings (seven steps for the starter task and six steps for the collector bowl task) served as an overall index of interjudge reliability. This reliability index could then be compared to the reliability of similar judgments made in previous studies in order to evaluate the suitability of the present study's expert raters' go/no-go ratings for use as criteria against which to assess the accuracy of subjects' subsequent ratings of the performances. High reliabilities would suggest that consistent, stable criterion target scores are being generated for the research. However, while high agreement for these expert ratings is important, it was felt that consensus judgments regarding the go/no-go ratings should be used since for a small number of task step performances, the final consensus differed from the initial majority rating. In other words, discussions revealed that a go or no-go opinion shared initially by the minority of panel members appeared to be more reasonable and justified than did the initial majority opinion.

Analysis of Predictor Variables and Development of Final Predictor Array. The predictor list presented in Figure 2 was examined and analyses conducted to reduce the number of variables for subsequent analyses involving correlating predictor measures with accuracy and other criteria. Specifically, most of the experience variables were highly intercorrelated, and so the two with the most meaning for the research (experience rating others and experience repairing jet engines) were retained. Self-rated

mechanical ability was dropped because the Mechanical Comprehension ASVAB subtest was seen as providing a better measure than the self-report. The high school GPA self-report was not included in the final predictor set due to potential problems with different grading policies at different schools. Regarding self-rating items dropped, number of hours slept the previous night did not have much variance, self-perceived tendency to criticize was quite similar to the evaluative tendency item, and self-reported accuracy in forming first impressions seemed less important upon reflection. Finally, three ACL components were formed based on past factor analyses results (Borman, 1979a), and, from the Post-Rating Questionnaire, the adherence to technical order scale had low variance and was eliminated. The reduced predictor set appears in Figure 4.

Jet Engine Experience: Experience as mechanic working on jet engine

Rating Experience: Number of times person has rated another mechanic

ASVAB AFQT: Armed Forces Qualification Test
[Arithmetic Reasoning + Mathematics Knowledge +
2(Word Knowledge + Paragraph Comprehension)]

ASVAB Mechanical Comprehension

Orientation Test (24 items)

Maze Test (24 items)

Detail Orientation

Rating Tendency: Tendency to attend to and evaluate the work of coworkers

Nitpick Tendency: Self-rated tendency to "nitpick"

ACL Achievement
(Achievement + Dominance + Self-Confidence - Succorance - Abasement)

ACL Social Closeness
(Nurturance + Affiliation + Deference - Autonomy - Aggression)

ACL Flexibility (Change - Order)

Figure 4. Reduced List of Predictors and Predictor Composites.

Coding and Examination of Reasons. As described above, respondents provided reasons for their go/no-go ratings. The instructions asked subjects to be especially careful to give reasons for their no-go ratings (i.e., what error or errors were observed); indeed, most respondents recorded reasons only for no-go ratings. Because the same reasons for performance errors often were expressed in a variety of ways, reasons needed to be grouped together to form categories and a coding scheme generated for each category. For example, on one step, the stated no-go reasons "wrong torque wrench technique" and "jerked torque wrench" appeared to represent two ways of conveying the same information and thus were judged to fit in a single category. Because similar reasons were provided in response to each step of the task across the six ratees, a coding scheme was devised by the second author for each performance step. Both authors used this coding system independently to categorize the responses of a subset of the subject raters. All discrepancies between coders were noted and resolved. As an example, Figure 5 provides the coding scheme for Step 3 of the starter installation task; the complete coding scheme for all steps appears in Appendix E.

---

0 No reason
1 General: did it wrong
2 Clamp in wrong position/not indexed properly
3 Clamp not seated/not seated enough/didn't use mallet
4 Tightened coupling nut before seating it/job done in wrong order
5 T-bolt positioned improperly in clamp/bolt installed improperly
6 Coupling damaged/bad clamp
7 Used washer on clamp
8 Didn't torque nut/tighten it/clamp not tight
9 Didn't keep tension on splines until torqued
10 Didn't support starter
11 Overtorqued coupling
12 Shouldn't use speed wrench
13 Didn't check for proper position of starter

Figure 5. Coding Scheme for Step 3 of Starter Installation Task.

Development of Rating Accuracy and Rater Style Scores. Several different rating accuracy and rater style scores were computed. The accuracy components and style variables are summarized in Figure 6. The appropriate scoring procedures were applied to each performance step. Scores for each task, as well as total scores, were computed by summing over the steps for each task and for both tasks combined. The reason for generating task-level scores for each accuracy and rating style measure was to allow computation of a reliability, or consistency, coefficient which indicates the stability of these scores across the two different tasks. Following the signal detection theory concepts outlined above, go/no-go ratings first were scored as a Hit, Miss, Correct Rejection or False Alarm, where a Hit represents a correct go rating, a Miss represents a false go rating, a Correct Rejection represents a correct no-go rating, and a False Alarm represents a false no-go rating. Each response was scored as a "Correct Response" if it was either a Correct Rejection or a Hit. Thus, the respondent received several rater accuracy component scores for the two tasks, separately and combined, including total number of Hits, Misses, False Alarms, Correct Rejections and Correct Responses.

Accuracy Indices	Rating Style Indices
Hit	Thoroughness
Miss	Criticalness
Correct Rejection	Nitpick
False Alarm	Generous
Rater Process Accuracy	
Differential Elevation	

Figure 6. Summary of Rater Accuracy and Rating Style Indices.

Recall that the above accuracy components have been characterized as _outcome_ accuracy indices and that the present research program intended to explore a measure of _process_ accuracy as well, using the reason-for-rating responses. Process accuracy was indexed by the congruence of the reason given by the respondent for a performance error on a step and the expert-provided reason for the error on that step. If the step was keyed "no-go," and if the reason given by the subject matched the theme of the expert-provided reason for a no-go rating, the process was scored as correct (Rater Process Accuracy = +1). In several cases, more than one reason was counted as correct. Rater Process Accuracy (RPA) was keyed as incorrect (i.e., -1) if the reason provided did not match the expert-provided reason _or_ if the no-go step was marked as a go (Miss), with no reason or an incorrect reason recorded. Respondents who provided vague reasons for no-go ratings (e.g., "Did it wrong") received a process score of 0 for those steps. Again, the RPA score was computed for each task separately and for the two tasks combined.

Experimental measures of "rating style" were also developed. First, a Thoroughness measure was computed by counting the total number of reasons (for go _and_ no-go ratings) offered by the subject rater across all steps of each task. Second, a measure designed to tap "Criticalness" was derived based on the total number of no-go ratings. Respondents received Thoroughness and Criticalness scores for each task and for the two tasks combined.

Third, an index was developed to measure the degree to which each respondent "nitpicked"; i.e., marked performance steps as no-go for relatively trivial reasons. Subject raters received one point on the Nitpick scale for each step on which they rated the step as a no-go and provided only a reason that was deemed a trivial criticism by the two

experimenters in consultation with the E-7 technical advisors. As with the other accuracy and rating style indices, this procedure led to scores for each task separately and for the two tasks combined. Fourth, an index termed "Generous" was calculated for each task, and the two tasks together. The Generous index reflected the number of steps on which a subject noted a criticism in the reason-for-rating blank <u>but</u> rated the step as a go.

Finally, a measure of the accuracy of the overall task evaluation ratings was computed using the mean of the experts' ratings on the 5-point scale as target scores. This measure was Differential Elevation--the degree of agreement between a rater's rating of each ratee on each task and the corresponding expert-provided rating, averaged across the two tasks and six ratees. It was computed using the following formula:

$$\sqrt{1/k \sum [(X_{i.} - X_{..}) - (T_{i.} - T_{..})]^2}$$

where k equals the number of ratees; $X_{i.}$ and $T_{i.}$ are mean rating and mean target scores for ratee j; and $X_{..}$ and $T_{..}$ are mean rating and mean target score over all ratees and dimensions. A measure of Overall Elevation was obtained by computing the mean overall rating for each task and for the two tasks combined. This is simply a measure of the average level of a rater's overall task performance evaluation ratings across all ratees.

III. RESULTS

<u>Expert Ratings of Videotaped Performance</u>

Expert go and no-go ratings of the videotaped performances, along with the final consensus ratings, are summarized in Appendix F. The consensus judgments were used as target criterion ratings against which to evaluate the observation accuracy of the 79 mechanic subject raters viewing the

videotapes.

As can be seen in Appendix F, interrater agreement was quite high for the five experts' independently derived go/no-go ratings. Intraclass correlations were computed for individual task steps using "1" for go and "0" for no-go ratings. These correlations ranged from .00 to 1.00, with a median of .94.[1] Discussions among experts, following independent ratings of each ratee on each task step, resolved differences of opinion in the 24 cases (of the 78 task steps) where there was less than perfect agreement on the part of the expert panel. In the vast majority of these 24 cases, the one or two experts who had ratings different from the majority, readily agreed with the majority opinion. In four cases, there was considerable discussion about the videotaped performances, and they were re-run several times before consensus could be reached. In five cases, the consensus judgment was different from the intended go or no-go rating scripted into the performances. Four of these involved the expert panel identifying a mistake or error that was not intended in the videotaping, and in the remaining case, the error built into the performance was simply not discernible in the videotaped depiction of that task step.

Thus, overall, agreement among panel members on the go and no-go ratings was quite high; where disagreement did occur, resolution of

---

[1]There was one task step (of the 13) with a very low intraclass correlation; this appeared to be more a function of restriction in range across the ratees in performance on the step (i.e., almost all no-go's) than of expert rater disagreement.

differences was reasonably easy and satisfactory. Further, the consensus judgments largely confirmed the intended ratings so that the pattern of go and no-go performances built into the target score matrix (see Table 1) was not substantially compromised.

Criteria

As described in the Methods section, the videotaped performances were presented to the raters in a different order for each session. This procedure eliminated any possibility of a consistent order effect in the ratings across sessions. Table 3 presents the means, standard deviations, inter-task consistencies and correlations between the 12 rater accuracy and style scores.

Observation and Evaluation Accuracy

Of 78 total performance steps across the two tasks, 33 were keyed no-go and 45 were keyed go according to the expert consensus judgments. On the average, raters made somewhat fewer no-go ratings (29.5) and thus more go ratings (48.5) than the actual performance base rates. In addition, Table 3 shows that of the 78 performance steps, the rater's go/no-go response was correct an average of 82% of the time (64 steps). Of the 33 steps on which the ratee made an expert-judged error, raters answered an average of 24 (73%) correctly. Similarly, on the 45 steps which experts judged as performed correctly, raters on average answered about 40 items (89%) correctly.

Table 3 also shows that the Number Correct score was less consistent across the two tasks than were the two scores that comprise this score--Hits and Correct Rejections. The inter-task correlations were .48 and .42 for Hits and Correct Rejections, respectively; this value was only .31 for Number Correct. This suggests that the Hits and Correct Rejection accuracy

Table 3. Means, Standard Deviations, Inter-Task Consistencies and Intercorrelations of 12 Rater Accuracy and Style Scores[a]

Mean	SD	Inter-task consistency[b]		Hits	Misses	Correct rejections	False alarms	Number correct	Criticalness (no-go's)	RPA	Nitpick	Generous	Thorough	Differential elevation	Average evaluation
39.50	4.21	.48/.65	Hits	1.0											
8.58	3.13	.40/.57	Misses	.41	1.0										
24.35	3.12	.42/.59	Correct Rejections	-.40	-1.0	1.0									
5.49	4.23	.48/.65	False Alarms	-1.0	-.41	.40	1.0								
63.71	4.23	.31/.47	Number Correct	.69	-.33	.34	-.69	1.0							
29.51	6.46	.57/.73	No-go's	-.87	-.78	.79	.87	-.24	1.0						
3.61	8.50	.43/.60	Process Accuracy	.57	-.23	.23	-.57	.77	-.23	1.0					
2.39	2.18	.21/.35	Nitpick	-.72	-.36	.36	.72	-.41	.67	-.44	1.0				
2.27	3.05	.66/.80	Generous	-.08	.24	-.26	.07	-.24	-.06	-.08	.15	1.0			
34.32	8.13	.61/.76	Thorough	-.69	-.52	.52	.68	-.22	.75	-.16	.50	.28	1.0		
.65	.18	.10/.18	Differential[c] Elevation	.30	.23	-.23	-.30	.11	-.33	.01	-.29	-.17	-.18	1.0	
2.69	.48	.66/.80	Average Evaluation	.53	.11	-.41	-.53	.20	-.57	.16	-.36	-.13	-.51	-.25	1.0

Note $p(.05) = .22 - .24$; $p(.01) = .29 - .31$.

[a] N = 66 to 78; one or two outliers per variable were removed because these data lay more than 3 standard deviations from the mean.

[b] Correlation between scores on starter installation task and scores on collector bowl task/Spearman-Brown-Adjusted Value to reflect reliability of mean, across-task scores.

[c] High Differential Elevation scores represented lower accuracy but the Differential Elevation scores have been reflected so that higher scores now indicate higher accuracy.

components are reasonably stable across tasks. Thus, if a rater had a high score on one of the components for one task, he/she was likely to have a high score on that component for the other task. The same pattern would be obtained for relatively low scorers. The comparably low reliability for the Number Correct measure may be due to the negative correlation between Hits and Correct Rejections (-.40), indicating accuracy on the steps keyed as no-go is negatively correlated with accuracy on the steps keyed as go.

Evaluation accuracy (Differential Elevation) and observation accuracy were moderately related when Hits defined observation accuracy ($r = .30$), but this relationship was lower when Number Correct represented observation accuracy ($r = .11$) and negative when Correct Rejections were used on the observation accuracy measure ($r = -.23$). However, these correlations may be depressed due to the low reliability of Differential Elevation; Differential Elevation calculated for the starter task correlated only .10 with Differential Elevation computed for the collector bowl task.

Rater Process Accuracy (RPA), the total number of correct reasons for no-go ratings, was highly correlated with overall observation accuracy; the correlation between RPA and Number Correct was .77. However, the relationships between RPA and the individual components of observation accuracy were smaller ($r = .57$ with Hits and $r = .23$ with Correct Rejections).

Other Criterion Scores

Both Nitpick and Generous scores had low means, indicating that on the average, very few raters marked a no-go for a relatively trivial reason (Nitpick) or marked a go while noting a performance error (Generous). Although somewhat unstable across the two tasks, the Nitpick score correlated highly with False Alarms ($r = .72$) and with total number of no-

gos (r = .67). Generous scores, on the other hand, were reasonably stable across the tasks (r = .66), but were relatively independent of the other criterion scores.

Thoroughness, the total number of reasons recorded across the 78 steps, was fairly stable across the two tasks (r = .61). By design, almost all no-go ratings were accompanied by one or more reasons. Thus, Thoroughness correlated highly with both number of no-gos (r = .75) and with number of False Alarms (r = .68).

Predictors

Correlations among selected predictor variables are shown in Table 4. Most of the relationships among the predictors are relatively low, although the two experience variables--jet experience and experience rating other mechanics--correlated .51, and the two ASVAB scores--AFQT and Mechanical Comprehension--correlated .58. The rest of the cognitive tests, the personality composites, and other self-report variables generally intercorrelate minimally. Thus, it appears that the predictor variable set measures a broad domain of individual differences with a relatively small degree of overlap.

Table 5 presents the correlations between selected predictors and the three process variables--starter and collector bowl task knowledge, motivation to make accurate ratings, and self-perception of how clearly details of the performances were seen. Respondents reporting confidence in how the tasks should be performed tended to have high AFQT, Orientation, and Task Effort scores. Individuals responding they tried hard to do the rating task well (motivation) were characterized by high detail orientation, a tendency to observe and evaluate the performance of coworkers, a tendency to nitpick, and, generally speaking, try hard on jobs

Table 4. Correlations Among Selected Predictors[a]

	Jet experience	Rating experience	AFQT	Mechanical comprehension	Orientation	Mazes	Detail orientation	Rating tendency	Sociability	Nitpick style	Task effort tendency	Achievement	Social closeness	Flexibility
Jet experience	1.00													
Experience rating other mechanics	.51	1.00												
ASVAB AFQT[b]	.17	-.18	1.00											
ASVAB Mechanical Comprehension[b]	.26	-.02	.58	1.00										
Orientation	-.22	-.35	.35	.17	1.00									
Mazes	-.42	-.35	.33	.35	.18	1.00								
Detail Orientation	.09	.07	.06	.22	-.06	.02	1.00							
Rating Tendency	.31	.13	.40	.23	-.06	-.08	.41	1.00						
Sociability	.12	-.03	-.03	-.27	.02	-.01	.07	-.08	1.00					
Nitpick Style	.13	.07	-.04	.06	-.07	-.05	.27	.42	-.06	1.00				
Task Effort Tendency	.10	.18	.05	-.03	.19	-.18	.31	.13	.04	.01	1.00			
Achievement	.03	.24	.13	.14	-.17	-.02	.09	.14	.15	.14	.11	1.00		
Social Closeness	.09	-.04	-.28	-.19	.04	-.06	.19	-.10	.23	-.02	.32	-.15	1.00	
Flexibility	-.45	-.18	-.13	-.20	.02	.26	-.21	-.24	-.09	-.10	-.32	-.12	-.31	1.00

Note $p(.05) = .22$; $p(0.1) = .29$ (.26 and .34 for $N = 54$).

[a] $N = 76$ to 79 except where noted.

[b] $N = 54$.

Table 5. Correlations Between Selected Predictors and 3 Process Variables[a]

	Task knowledge	Motivation	Ability to see
Jet experience	-.06	.10	-.14
Experience rating other mechanics	-.02	.03	.08
ASVAB AFQT[b]	.20	.11	-.17
ASVAB Mechanical[b] Comprehension	-.07	.11	-.21
Orientation	.24	.08	.00
Mazes	-.05	-.14	-.23
Detail Orientation	.17	.22	-.07
Rating Tendency	.14	.30	.10
Sociability	.09	-.11	-.06
Nitpick Style	.10	.24	-.07
Task Effort	.29	.21	.05
Achievement	.15	-.00	-.04
Social Closeness	-.03	.20	.06
Flexibility	.08	-.20	.00

[a]N = 76 to 79 except where noted.

[b]N = 54.

* $p < .05$.
** $p < .01$.

and tasks. Correlations with self-perceived ability to see considerable detail in the videotaped performances tended to be low and in several cases were counterintuitive. For example, scores on the Maze test and ASVAB Mechanical Comprehension both correlated negatively with this self-report measure (-.23 and -.21, respectively).

Relationships Between Predictors and Rating Criteria

Correlations between the reduced set of predictors and the 12 rating accuracy and rater style variables appear in Table 6. This table reveals several somewhat surprising findings. First, experience working on a jet engine and experience rating other mechanics correlated negatively with Hits ($r = -.42$ and $-.38$, respectively). Thus, more experienced raters tended to have fewer correct go ratings than did the less experienced raters. Raters with relatively more jet engine maintenance experience also tended to demonstrate lower Process Accuracy (RPA $r = -.17$).

On the other hand, these experience variables correlated in the expected direction with Correct Rejections ($r = .26$ for experience working on jet engines and .24 for experience rating other mechanics). It appears from these results that more experienced raters tended to be more critical than less experienced raters, identifying performance errors even when these errors were not programmed into the videotapes. This is further reflected in the tendency for more experienced raters to give the task performances overall evaluation ratings that were lower than those given by the less experienced raters, to rate more steps as no-go, and to nitpick more often than their less experienced counterparts.

In contrast to the experience variables, AFQT scores correlated positively (although not significantly) with Number Correct ($r = .22$) and Correct Rejections ($r = .25$). Individuals with high AFQT scores also

Table 6. Correlations Between Selected Predictors and 12 Criteria[a]

	Hits	Misses	Correct rejections	False alarms	Number correct	Criticalness (no-go's)	RPA	Nitpick	Generous	Thorough	Differential[c,d] elevation	Average[c] evaluation
Experience with jet engine	-.42**	-.27*	.26*	.42**	-.20	.38	-.17	.33**	.11	.46**	-.27*	-.27*
Experience rating other mechanics	-.38**	-.25*	.24*	.38**	-.17	.30**	-.15	.29**	.14	.36**	-.32**	-.30**
AFQT[b]	.04	-.23	.25	-.04	.22	.10	.24	-.11	-.10	.22	.08	.05
ASVAB Mechanical[b] Comprehension	-.10	-.14	.17	.10	.02	.16	.05	.01	-.07	.27*	.21	-.14
Orientation	.26*	-.06	.07	-.26*	.29**	-.09	.43**	-.27*	.01	-.09	.04	.11
Mazes	.19	.07	-.07	-.19	.13	-.12	.18	-.03	.09	-.27*	-.04	-.03
Detail Orientation	-.12	-.18	.18	.12	.02	.20	.11	.10	-.05	.28*	.09	-.06
Rating Tendency	-.25*	-.22*	.21	.25*	-.08	.26*	-.02	.26*	.16	.34**	.11	-.14
Sociability	-.12	.03	-.05	.12	-.12	.10	-.01	.20	.12	.06	-.16	-.03
Nitpick Style	-.29**	-.22*	.22*	.30**	-.07	.38**	-.05	.27*	-.07	.26*	.02	-.26*
Task Effort	-.20	-.14	.14	.20	-.02	.21	-.01	.05	.08	.29**	-.16	-.16
Achievement	-.07	-.07	.06	.07	.01	.10	-.05	-.02	.00	.13	-.09	-.23*
Social Closeness	-.18	-.07	.07	.18	-.11	.17	-.11	.18	-.01	.15	-.26*	.08
Flexibility	.26*	.23*	-.23*	-.26*	.09	-.29**	.13	-.18	.08	-.24*	.10	.04

[a] N = 76 to 79 except where noted.
[b] N = 54.
[c] N = 68.
[d] Differential Elevation scores have been reflected so that higher scores indicate higher accuracy.
*p < .05.
**p < .01.

tended to have high Process Accuracy (r = .24) and high Thoroughness (r = .22). Correlations between Mechanical Comprehension and the rater accuracy and style scores were nonsignificant except for the correlation with Thoroughness (r = .27).

Orientation test scores were correlated substantially with Hits (r = .26) and Number Correct (r = .29), as well as with RPA (r = .43). In contrast, scores on the Maze test correlated only .19 with Hits, .13 with Number Correct and .18 with RPA.

Four of the background predictors had very similar patterns of correlations with the rating accuracy and rater style variables. Detail Orientation, Rating Tendency, Nitpick Style, and Task Effort all were related to the Criticalness of the rater, correlating .20, .26, .38 and .21, respectively, with the number of no-go's. These predictors were all positively correlated with the number of Correct Rejections but negatively correlated with the number of Hits.

Although most of the relationships between the three Adjective Checklist composites and the rating accuracy and rater style variables were small and nonsignificant, these results suggest that more flexible raters are less critical (they rate fewer steps as no-go's; r = -.29) and make fewer False Alarms (r = -.26). Also, individuals with higher Achievement Orientation tend to give the ratees lower overall task evaluations (r = -.23), and Social Closeness correlates negatively with evaluation accuracy (i.e., Differential Elevation; r = -.26).

Table 7 contains correlations between the three process variables and the 12 criteria. From this table it appears that those respondents who believed they knew how the tasks should be performed tended to have more total Number Correct go/no-go ratings than did those who were less confident in their task knowledge (r = .28). A similar patte'n of results

Table 7. Correlations Between 3 Process Variables and 12 Criteria[a]

	Task knowledge	Motivation	Ability to see
Hits	.15	-.23*	.28*
Misses	-.13	-.09	.03
Correct Rejections	.13	.09	-.03
False Alarms	-.15	.23*	-.28*
Number Correct	.28*	-.14	.27*
Criticalness	-.01	.25*	-.20
RPA	.29**	-.07	.12
Nitpick	-.01	.19	-.20
Generous	-.08	.16	-.11
Thorough[b]	.10	.31**	-.21
Differential Elevation[b]	-.15	-.02	-.14
Average Evaluation	.06	-.20	.30**

[a]N = 76 to 79 except where noted.

[b]N = 68.

* $p < .05$.
** $p < .01$.

appears for self-perceived ability to see details of the videotaped performances. Respondents who believed they could see the performances well tended to have more Hits ($r = .28$) and more total Number Correct ($r = .27$). Also, self-rated ability to see details in the videotapes correlated positively with Average Evaluation ($r = .30$). In contrast, self-rated motivation to do well on the task correlated most highly with the rater style measures Criticalness and Thorough ($r = .25$ and $r = .31$, respectively), rather than with the accuracy components.

IV. DISCUSSION

This section focuses on the usefulness of the videotaped performances as stimulus materials, relationships among the observation accuracy components, relevance of signal detection theory to studying observation accuracy, usefulness of rating styles as a concept, and relationships between rater individual differences and accuracy component scores.

Usefulness of the Videotape Performances for Personnel Research Applications

The videotapes of six scripted performances on the two installation tasks should provide useful stimulus materials for future research applications. Considerable work was done to make the performances realistic, both in the filming and in configuring videotaped performance levels to correspond to actual performance levels for first-term mechanics. As a result, the skills and abilities necessary to provide accurate ratings for the videotapes should be the same skills and abilities required for making actual Walk-Through Performance Test hands-on ratings.

Thus, the videotapes and accompanying rating scales and technical orders should be useful for selecting the most accurate WTPT scorers or training these scorers to provide even more accurate ratings. For example, raters could use the tapes to practice making observations (go/no-go

ratings) and evaluations (overall ratings for each ratee on each task). Feedback on how their rating performance differed from the target scores, along with the expert-provided reasons for errors where there are scripted errors, could greatly aid in the rater training process. The videotapes also provide an instrument for evaluating the impact of other training programs. Trainees could rate the videotaped performances before and after some experimental training treatment (or rate half of the performances before and half after training). Differences between pre- and post-training performance could be used as a measure of the impact of training on relevant accuracy components.

Issues Concerning Rating Accuracy Criteria

Reliability

An interesting and promising finding in the research program was that scores on most of the accuracy components are reasonably consistent across the two tasks. Hits and Correct Rejections both have reliabilities near .60, as do the respective reflections of these components, Misses and False Alarms. This suggests that accuracy components have some validity in the sense that raters who are accurate on a component for one part of the observation task will likely be reasonably accurate on that component for other parts of the observation task. Having reliable accuracy component scores also means that interpretable relationships with other variables are possible, as was found in this study. Finally, these reliability results are consistent with previous findings (Borman, 1977) that showed halo and restriction-in-range rating errors, along with differential accuracy, to be reasonably stable phenomena.

In addition to the observation accuracy components, Rating Process Accuracy (RPA) was found to be stable across tasks, as were the rater style

measures Thoroughness and Generous. The Nitpicking style variable and the Differential Elevation accuracy index had lower reliabilities.

Correlations Among Accuracy Criteria

An important result in the study was the negative correlation between the two accuracy components Hits and Correct Rejections ($r = -.40$). On the one hand, this finding clearly demonstrates the value of considering the accuracy components separately. If Hits and Correct Rejections had been included only as part of the Number Correct composite, the differing patterns of individual difference correlates for these two measures would have been masked. However, if Hits and Correct Rejections are thought of as two underlying components of an overall accuracy construct, the negative correlation between them is bothersome and perplexing.

Why do we find this negative relationship? The answer may be that rater criticalness is more important than rater accuracy as a rater behavioral characteristic underlying the Hit and Correct Rejection outcome accuracy components. As a rater becomes more critical, he or she will tend to have more Correct Rejections, but fewer Hits. For example, those raters who had many Hits tended to be less critical than those who had many Correct Rejections.

The correlations with the global Number Correct measure in Table 6 suggest that the most accurate raters, overall, tended to be those with high cognitive ability (AFQT score), high spatial ability (Orientation Test scores), but relatively little experience in rating other mechanics and working on jet engines. However, the results look quite different when Correct Rejections and Hits are used individually as indices of rater accuracy. Raters having the most Correct Rejections tended to have high cognitive and spatial ability, but they also tended to be more experienced

as raters and as mechanics. Also, they had a tendency to be relatively evaluative about others' job performance, to "nitpick," and to be less flexible than those subjects who had fewer Correct Rejections.

Perhaps the most surprising finding in the study was that jet engine repair and performance rating experiences are negatively correlated with Hits. As mentioned, the reason for this result may lie simply in the tendency for the more experienced raters to be more critical. One explanation for this is experienced raters are more aware (than their less experienced counterparts) of the importance of detecting errors in all work on a jet engine. Within this setting, Misses are definitely to be avoided; an improperly installed part, for example, could lead to a very expensive (or even lethal) mistake. Thus, Correct Rejections take on greater importance than Hits (Hedge, Dickinson, & Bierstedt, 1988). The experienced mechanic knows this, and when a marginally acceptable performance is observed, he/she is more likely than the less experienced mechanic to grade the performance as a no-go.

This argument regarding the more experienced subject is indirectly supported by examining no-go base rates of all the subjects against the same base rates reflected in the consensus judgments made by very experienced mechanics. The subjects, on the average, made fewer no-go ratings than was evident in the consensus judgments (29.5 versus 33). Taking into account the substantial positive correlation between experience level and number of no-go's in the subject sample, it is likely that the less experienced raters provided, on the average, considerably fewer no-go's than were indicated by the consensus judgments and that the more experienced raters (again, on the average) had <u>about the same</u> base rate of no-go responses as did the expert judges.

Evaluation-Observation Accuracy Relationships

The present study also afforded an opportunity to examine relationships between evaluation accuracy (Differential Elevation) and components of observation accuracy. In general, these correlations were lower than those presented by Murphy et al. (1982). The highest such relationship in this study was .30 between Differential Elevation and Correct Responses. In contrast, Murphy et al. found some observation-evaluation accuracy correlations in the .40 range and higher. The relatively low reliability of the Differential Elevation index could have limited the magnitude of the relationships found in the present study.

The Concept of Process Accuracy and the Rating Process Accuracy Measure

In the present research, it was argued that a possible measure of process accuracy, beyond Lord's (1985) suggestion of using Hits and False Alarms, was a Rating Process Accuracy (RPA) index requiring both correct go and no-go ratings <u>and</u> the correct reason(s) for no-go marks. Reliability of the measure was reasonably high ($r = .65$). In addition, the pattern of correlations with predictor variables was sufficiently different from the patterns found for other rating criterion measures that RPA may deserve further exploration as a process accuracy index.

The Concept of Rater Styles

Three rater style variables were identified and explored in the research: Nitpick, Generous, and Thorough. The rationale for this interest in rater style was that although accuracy seems to be the most important "bottom-line" outcome criterion to describe rating behavior, knowledge about such behavior could be augmented by examining rater styles and their individual difference correlates.

Two of the three style measures were very reliable and the third

(Nitpick) was less consistent across the two tasks ($r = .24$, task x task correlation). The Thorough measure correlated substantially with experience in jet engine maintenance ($r = .46$) and with Mechanical Comprehension ($r = .27$). It also correlated significantly with the two motivation measures, self-reports of a general tendency to try hard on tasks ($r = .29$) and of effort put forth on this rating task ($r = .31$). Although not very reliable, the Nitpick variable was useful in helping to understand the predictor-accuracy component correlations. Nitpick correlated .33 with jet engine mechanic experience, suggesting that the substantial negative correlation between experience and Hits, while being correlated positively with False Alarms, arose, in part, because the more experienced subject raters were identifying relatively insignificant reasons for some of their no-go ratings.

Also, certain relationships obtained in the study provide some evidence for these measures' construct validity. Self-perceived Nitpick style correlated .27, for example, with the independently derived rating behavior Nitpick measure. Flexibility correlated -.29 with Nitpick style (criticalness). Likewise, and as mentioned, the motivation self-reports correlated significantly with the Thoroughness style measure ($r = .31$); detail orientation self-report also correlated substantially ($r = .28$) with the Thoroughness index.

Future Research

The videotapes themselves, along with supporting rating scale and technical order materials, can be used in future personnel research efforts. Because of the numbers of variables and limited size of the sample, more complex statistical analyses regarding predictor-criterion relationships were not feasible within the present research. However,

correlations were sufficiently high between rater individual differences and process variables such as Task Knowledge and Motivation to Rate, and between the process variables and accuracy components, that causal analyses using some of the same measures might profitably proceed with larger samples. In this way, links may be more precisely defined among measures in these different variable sets.

References

Baker, E., & Schuck, J. R. (1975). Theoretical note: Use of signal detection theory to clarify problems of evaluating performance in industry. Organizational Behavior and Human Performance, 13, 307-317.

Borman, W. (1977). Consistency of rating accuracy and rater errors in the judgment of human performance. Organizational Behavior and Human Performance, 20, 238-252.

Borman, W. (1978). Exploring upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 63, 135-144.

Borman, W. (1979a). Individual differences correlates of accuracy in evaluating others' performance effectiveness. Applied Psychological Measurement, 3, 103-115.

Borman, W. (1979b). Format and training effects on rating accuracy and rater error. Journal of Applied Psychology, 64, 410-421.

Borman, W., & Cascio, W. (1982). Consistency of rating accuracy and rater errors in a sample of hospital nurses and administrators. Unpublished manuscript, Minneapolis, MN.

Gough, H., & Heilbrun, A. (1965). The Adjective Check List Manual. Palo Alto, CA: Consulting Psychologists Press.

Green, D., & Swets, J. (1966). Signal detection theory and psychophysics. New York: Wiley.

Hedge, J. W., Dickinson, T. L., & Bierstedt, S. A. (1988, July). The use of videotape technology to train administrators of Walk-Through Performance Testing (AFHRL-TP-87-71, AD-A195 944). Brooks Air Force Base, TX: Training Systems Division, Air Force Human Resources Laboratory.

Hedge, J. W., & Teachout, M. S. (1986, November). Job performance measurement: A systematic program of research and development (AFHRL-TP-86-37, AD-A174 175). Brooks Air Force Base, TX: Training Systems Division, Air Force Human Resources Laboratory.

Kavanagh, M., MacKinney, A., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 75, 34-49.

Kraiger, K. (1985). Analysis of relationships among self, peer, and supervisory ratings of performance (Contract No. F49620-85-C-0013). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Lord, R. (1985). Accuracy in behavioral measurement: An alternate definition based on raters' cognitive schema and signal detection theory. Journal of Applied Psychology, 70, 66-71.

Murphy, K., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. (1982). Relationships between observational accuracy and accuracy in evaluating performance. Journal of Applied Psychology, 67, 320-325.

Peterson, N. G. (Ed.). (1987). Development and field test of the trial battery for Project A (Technical Report 739). Alexandria, VA: U.S. Army Research Institute.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Thornton, G., & Zorich, S. (1980). Training to improve observer accuracy. Journal of Applied Psychology, 65, 351-354.

Vance, R., MacCallum, R. E., Coovert, M. S., & Hedge, J. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. Journal of Applied Psychology, 73, 74-80.

APPENDIX A:

DESCRIPTIONS OF TASK STEPS AND

SCRIPTED (INTENDED) TARGET SCORE MATRICES

INSTALL STARTER

Step 1: Apply petroleum or grease to spline.

Step 2: Hang clamp.

Step 3: Index starter in appropriate position.

Step 4: Tighten and seat V-band clamp.

Step 5: Torque.

Step 6: Install locking device on V-band clamp.

Step 7: Make and safety-wire electrical connection.

INSTALL BLEED AIR SYSTEM COLLECTOR BOWL

Step 1: Apply petro to clamp assembly.

Step 2: Index flapper valve assembly and install gaskets.

Step 3: Position collector bowl coupling.

Step 4: Install all gaskets and ducts properly.

Step 5: Use straight edge to position ducts 2 1/2 inches from top of rim.

Step 6: Install backup nuts on stub duct and safety-wire.

Intended Target Score Matrix for Install Starter

Steps	Ratees 1	2	3	4	5	6	% Go
1	G	G	G	N	N	G	67
2	G	G	N	G	G	G	83
3	G	G	N	G	G	G	83
4	G	N	N	G	N	N	33
5	G	G	N	N	N	G	50
6	G	G	G	G	N	G	83
7	G	N	G	N	G	G	67

G = Go.

N = No-Go.

Intended Target Score Matrix for Install Bleed Air System Collector Bowl

	Ratees						
Steps	1	2	3	4	5	6	% Go
1	G	G	G	N	N	G	67
2	G	N	N	N	G	G	50
3	G	G	G	N	G	G	83
4	G	N	G	G	N	G	67
5	G	N	G	N	N	G	50
6	G	N	N	N	N	G	33

G = Go.

N = No-Go.

APPENDIX B:

PERFORMANCE RATING BOOKLETS AND

TECHNICAL ORDERS FOR

STARTER AND COLLECTOR BOWL TASKS

Performance Rating Booklet

Performance Rating Instructions

Go/No-Go Ratings

For each step that a mechanic on the videotape performs, record on the Performance Rating Sheet whether or not the mechanic performed the step correctly. After each performer completes each step of the task, check "go" or "no-go" on the appropriate line of the rating sheet corresponding to that step of the task. A "go" means the performer did the step correctly, and a "no-go" means he made one or more mistakes. Remember to wait until the step is complete before you make your rating. After you check "go" or "no-go," explain briefly why you rated the performance as you did. Use the line labeled "reason" to provide your reason. If on a particular step you rated a performer "go" and you simply noticed no flaws or problems in the performance, you may write "nothing wrong." However, there may be several performances you mark as "go" where you will want to give some other reason besides "nothing wrong." When you mark a performance "no-go" you should always record your reason.

Overall Performance Ratings

When the mechanic has completed the entire task, rate on a scale of 1 to 5 the overall quality of the mechanic's performance. Use the 1-5 scale displayed at the bottom of each page. This scale includes a description of high and low performance and looks like this:

Low				High
1	2	3	4	5
Mechanic performed the task incorrectly, using improper tools, materials or methods, and completing the job inefficiently or with poor technique.			Mechanic completed the job correctly and efficiently, using the appropriate tools and methods, and showing good technique.	

The statement at the left describes the performance of a person who should receive a rating of 1 or 2; the statement at the right describes the performance of a person who should receive a rating of a 4 or 5. A 3 rating would indicate performance at a level roughly mid-way between the high and low statements.

For example, assume that a performer completed the starter assembly installation perfectly, and completed Steps 1-4 and Step 6 of the collector bowl task correctly, but made errors on Steps 5 and 7 while installing the collector bowl. The evaluator's rating sheet for this mechanic should look like the one on the next page (except that the reason lines have been left blank for the example).

Performance Rating Sheet (Example)

Starter Assembly Installation

Step	Description	Go	No-Go
Step 1:	Apply grease to spline	✓	
Reason:			
Step 2:	Hang clamp	✓	
Reason:			
Step 3:	Index starter in appropriate position	✓	
Reason:			
Step 4:	Seat and tighten V-band clamp	✓	
Reason:			
Step 5:	Torque coupling nut	✓	
Reason:			
Step 6:	Install and torque coupling safety nut	✓	
Reason:			
Step 7:	Make and safety-wire electrical connection	✓	
Reason:			

Overall rating 5

Rating Scale:

Low				High
1	2	3	4	5
Mechanic performed the task incorrectly, using improper tools, materials or methods, and completing the job inefficiently or with poor technique.			Mechanic completed the job correctly and efficiently, using the appropriate tools and methods, and showing good technique.	

Performance Rating Sheet (Example)

Collector Bowl Installation

Step	Description	Go	No-Go
Step 1:	Apply petroleum to clamp assembly	✓	
Reason:			
Step 2:	Index flapper valve assembly and install gaskets	✓	
Reason:			
Step 3:	Connect Y duct to collector bowl	✓	
Reason:			
Step 4:	Connect collector bowl to stub duct	✓	
Reason:			
Step 5:	Use straight edge to position ducts 2 1/2 inches from top of rim		✓
Reason:			
Step 6:	Install and torque backup nuts on stub duct and safety-wire	✓	
Reason:			
Step 7:	Torque nuts		✓
Reason:			

Overall rating 3

Rating Scale:

Low				High
1	2	3	4	5
Mechanic performed the task incorrectly, using improper tools, materials or methods, and completing tne job inefficiently or with poor technique.			Mechanic completed the job correctly and efficiently, using the appropriate tools and methods, and showing good technique.	

Note that the evaluator checked "go" for all the steps in the starter installation task, along with the first four and Step 6 for the collector bowl task, but checks "no-go" for Steps 5 and 7 of the collector bowl task.* Notice also that the evaluator made a rating of 5 for the starter installation performance, and a rating of 3 for the collector bowl performance. These overall task performance ratings were made taking into account the overall effectiveness of the mechanic on the respective tasks.

There are several assumptions or rules that we would like you to keep in mind as you evaluate the performances in the videotape:

1. The engine is being worked on in the shop, not on the aircraft.

2. The mechanics have been provided with the appropriate parts; in other words, you should not give a "no-go" simply because you would have used a different part.

3. Evaluate the performance you can see, not what you cannot see (assume what you do not see is performed correctly).

4. The mechanic may receive help from a second mechanic. Evaluate the performance of the primary mechanic.

5. When judging whether a mechanic working on a step should receive a "go" or "no-go," evaluate only the performance during that step; in other words, do not mark "no-go" because a previous step was performed incorrectly and/or was not corrected.

6. All mechanics on the tape are in their first term.

7. Regarding the go/no-go ratings, use your best judgment about whether or not the mechanic completes each task step properly. The technical order instructions should provide good guidance, but your own judgment will be important. Remember, mistakes in technique count as "no-go," as do errors in the choice of tools/materials and non-adherence to the technical order.

In all, you will be viewing the performance of six mechanics. Each mechanic will first perform the starter task and then the collector bowl task.

Before beginning the rating task, we would like to make an important request. It is absolutely critical that you remain quiet during the rating session. Do not say or do anything that might hint to the other evaluators that a performer has done something right or wrong. Again, please do not make any noise or give any hints to the other evaluators in the room while the videotape is playing. Now turn to the next page, entitled "Performance Rating Sheet 1A," and prepare to assess the work performance on the video screen. The administrator will answer any questions now.

* Author note: Although Step 7 of the Collector Bowl task was dropped to reduce the length of the videotapes, these instructions continued as is. It was explained to subjects that Step 7 had been eliminated.

Performance Rating Sheet (Prototype)

Starter Assembly Installation

Step	Description	Go	No-Go
Step 1:	Apply grease to spline	____	____
Reason:	____________________		
Step 2:	Hang clamp	____	____
Reason:	____________________		
Step 3:	Index starter in appropriate position	____	____
Reason:	____________________		
Step 4:	Seat and tighten V-band clamp	____	____
Reason:	____________________		
Step 5:	Torque coupling nut	____	____
Reason:	____________________		
Step 6:	Install and torque coupling safety nut	____	____
Reason:	____________________		
Step 7:	Make and safety-wire electrical connection	____	____
Reason:	____________________		

Overall rating ____

Rating Scale:

Low				High
1	2	3	4	5
Mechanic performed the task incorrectly, using improper tools, materials or methods, and completing the job inefficiently or with poor technique.			Mechanic completed the job correctly and efficiently, using the appropriate tools and methods, and showing good technique.	

Performance Rating Sheet (Prototype)

Collector Bowl Installation

Step	Description	Go	No-Go
Step 1:	Apply petroleum to clamp assembly	____	____
Reason:	____________________		
Step 2:	Index flapper valve assembly and install gaskets	____	____
Reason:	____________________		
Step 3:	Connect Y duct to collector bowl	____	____
Reason:	____________________		
Step 4:	Connect collector bowl to stub duct	____	____
Reason:	____________________		
Step 5:	Use straight edge to position ducts 2 1/2 inches from top of rim	____	____
Reason:	____________________		
Step 6:	Install and torque backup nuts on stub duct and safety-wire	____	____
Reason:	____________________		

Overall rating ____

Rating Scale:

Low				High
1	2	3	4	5
Mechanic performed the task incorrectly, using improper tools, materials or methods, and completing the job inefficiently or with poor technique.			Mechanic completed the job correctly and efficiently, using the appropriate tools and methods, and showing good technique.	

Technical Orders

for

Starter Assembly and Collector Bowl

Installation

Starter Assembly Instructions

(Diagrams appear on the next page)

1. Apply grease to all spline teeth on starter output shaft (3).

2. Place coupling (1) on adapter (2) and properly lock coupling latch to hold coupling on adapter.

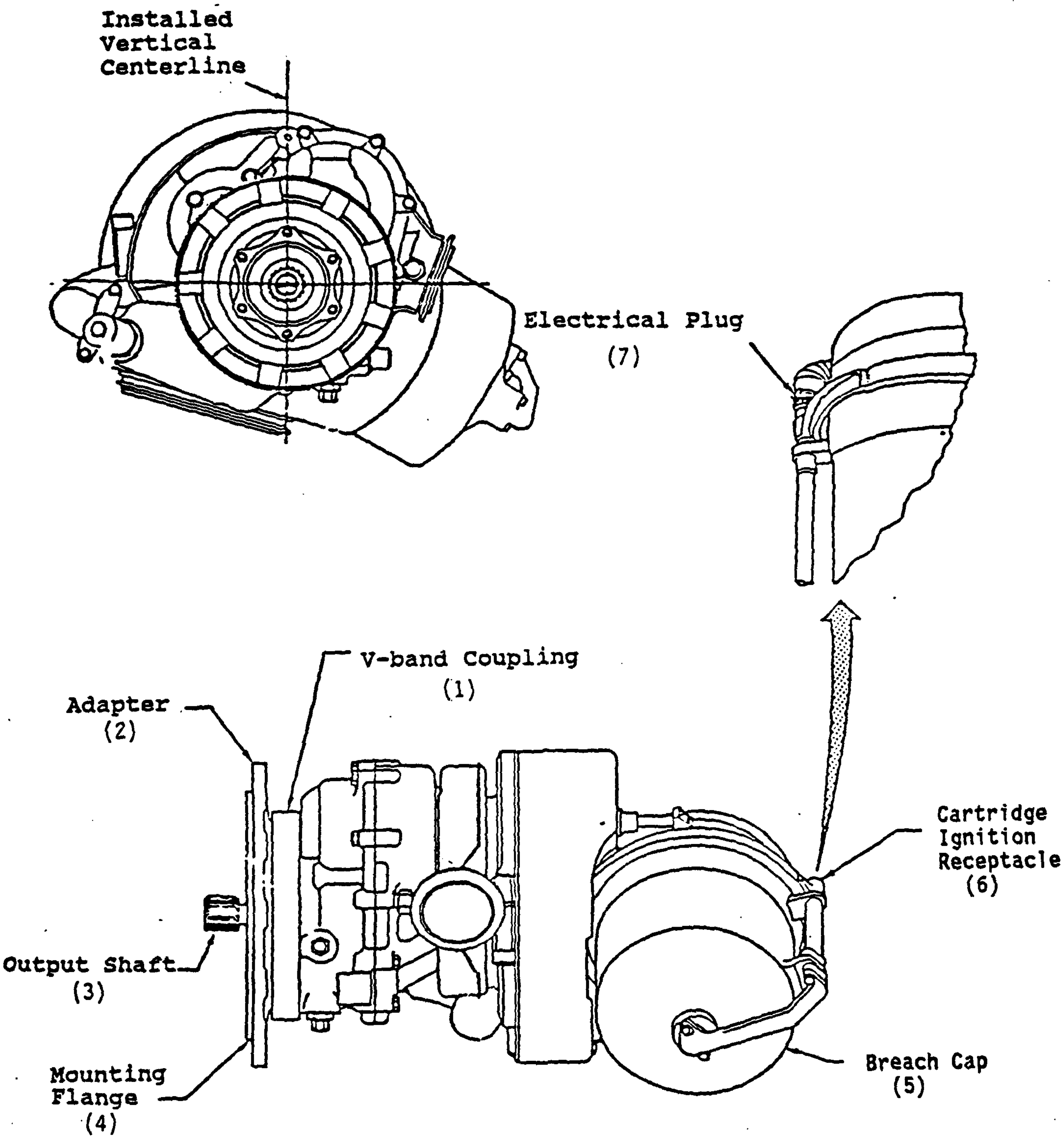
3. Raise starter into position and engage starter output shaft (3) with transfer gearbox splines. Rotate starter until breech chamber is at 8 o'clock position and position starter forward, working starter flange (4) under locking edge of coupling and thread coupling nut.

4. Ensure that coupling (1) is positioned so T-bolt is at 3 or 9 o'clock position, seat coupling and tighten coupling nut.

5. Torque existing coupling nut.

6. Install safety nut on coupling (1) and torque.

7. Connect the electrical plug (7) to the cartridge ignition receptacle (6) (on top of the starter), finger-tighten the plug, and safety-wire it with the proper lockwire.



Sundstrand Starter
Cartridge/Pneumatic Starter Installation

Collector Bowl Installation Instructions

(Diagrams appear on the next page.)

1. Apply light coat of Petrolatum to inner surfaces of both halves of clamp assembly (6).

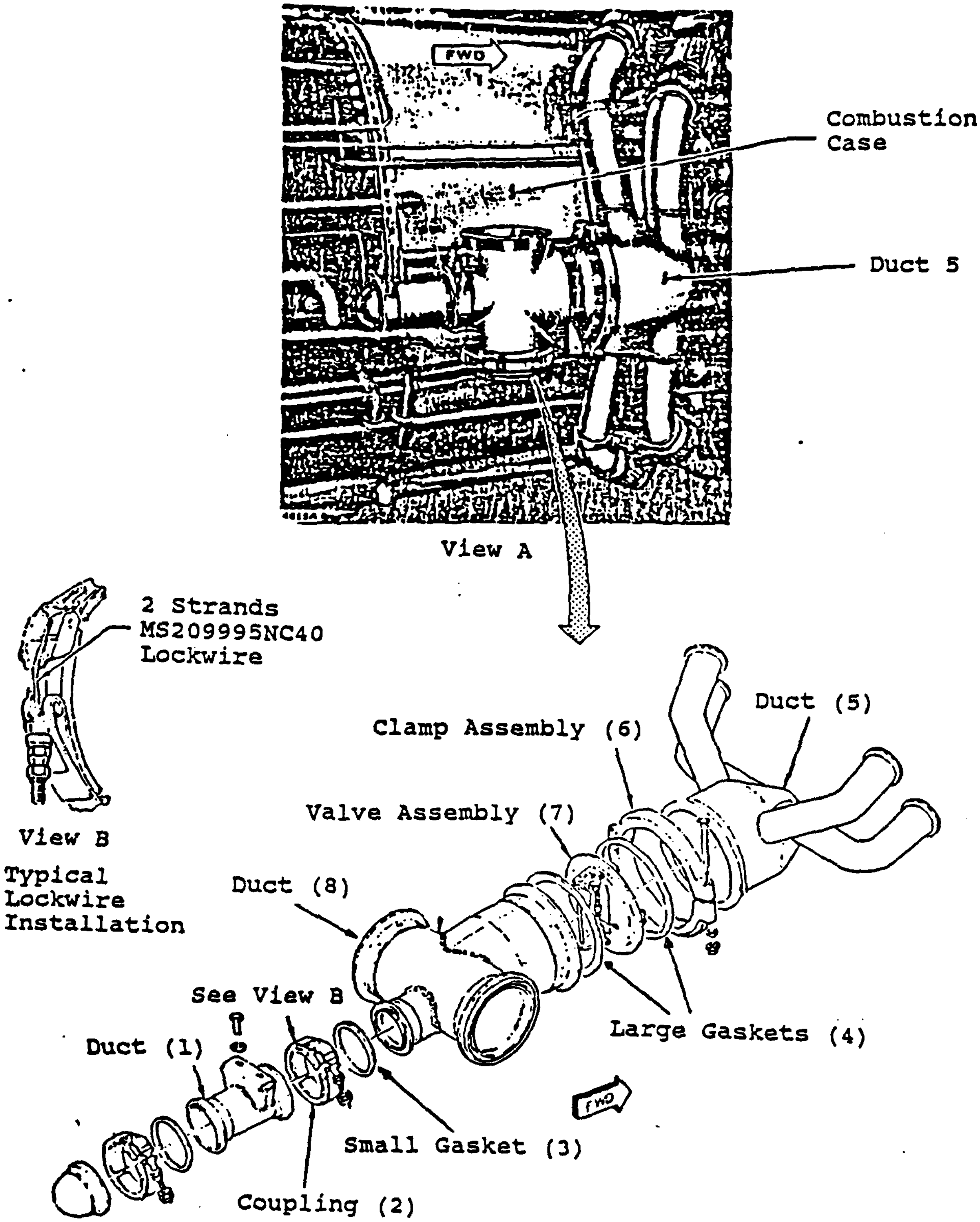
2. Install large gaskets (4) and index valve assembly so that square end of hinge pin points up or flow arrow points toward the small port of the duct (8).

3. Assemble duct (8) to duct (5) with two large gaskets (4), valve assembly (7), clamp assembly (6), and bracket assembly clamp. Install two bolts, head side up, with washers and nuts to clamp assembly. Use one washer under each nut. Do not torque at this time.

4. Place small gasket (3) between ducts (1 and 8) and join ducts using coupling (2). Rotate coupling to upper portion of the duct between the 3 and 9 o'clock positions, and thread nut.

5. Hold ruler against the open side of the duct (8) and adjust the duct to obtain a 2 1/2-inch space from the top of the duct flange to the combustion case.

6. Install and torque backup nuts on coupling (2) and safety-wire the coupling with double-strand lockwire. (See view B.)



Compressor Bleed Air Manifold Installation

APPENDIX C:

FINAL CONSENSUS RATINGS AND REASONS FOR NO-GO'S

INSTALL STARTER

Ratee 1

Step 1: Go

Step 2: Go

Step 3: No-Go - Coupling nut not threaded.

Step 4: Go

Step 5: Go

Step 6: Go

Step 7: Go

INSTALL STARTER

Ratee 2

Step 1: Go

Step 2: Go

Step 3: No-Go - Starter not in position and indexed wrong; coupling nut not threaded.

Step 4: No-Go - Install clamp at 6:00 position instead of either 3:00 or 9:00 position.

Step 5: Go

Step 6: Go

Step 7: No-Go - Put safety wire on backwards.

INSTALL STARTER

Ratee 3

Step 1: Go

Step 2: No-Go - Place coupling on starter and lock coupling latch to hold coupling on starter.

Step 3: No-Go - Position starter at 10:00 positionn instead of 8:00 position - latter is correct; coupling nut not threaded.

Step 4: No-Go - Improperly connect quick release portion of clamp to T-bolt.

Step 5: No-Go - Hold torque wrench wrong; i.e., hold it above the handle.

Step 6: Go

Step 7: Go

INSTALL STARTER

Ratee 4

Step 1: No-Go - Lubricate the transfer gear box splines.

Step 2: Go

Step 3: Go

Step 4: Go

Step 5: No-Go - Jerk torque wrench as opposed to using a smooth, even motion.

Step 6: Go

Step 7: No-Go - Leave the cannon plug really loose--must be finger tight.

INSTALL STARTER

Ratee 5

Step 1: No-Go - Lubricate only some of the spline teeth area.

Step 2: Go

Step 3: Go

Step 4: No-Go - Don't seat clamp; i.e., does not tap into it.

Step 5: No-Go - Torque with one hand only instead of two.

Step 6: No-Go - Tighten but don't torque safety nut.

Step 7: Go

INSTALL STARTER

Ratee 6

Step 1: Go

Step 2: Go

Step 3: Go

Step 4: No-Go - Install clamp without seating it; i.e., does not tap it.

Step 5: Go

Step 6: Go

Step 7: Go

INSTALL BLEED AIR SYSTEM COLLECTOR BOWL

Ratee 1

Step 1: Go

Step 2: No-Go - Left out gasket.

Step 3: Go

Step 4: Go

Step 5: Go

Step 6: No-Go - Didn't torque first nut.

INSTALL BLEED AIR SYSTEM COLLECTOR BOWL

Ratee 2

Step 1: Go

Step 2: No-Go - Misposition/misalign valve assembly.

Step 3: Go

Step 4: No-Go - Don't use gasket.

Step 5: No-Go - Have measure be 2 1/2 inches to bottom of rim as opposed to top of rim.

Step 6: No-Go - Install only one nut.

INSTALL BLEED AIR SYSTEM COLLECTOR BOWL

Ratee 3

Step 1: Go

Step 2: Go

Step 3: Go

Step 4: Go

Step 5: Go

Step 6: No-Go - Clamp improperly seated; therefore, had difficulty tightening nut.

INSTALL BLEED AIR SYSTEM COLLECTOR BOWL

Ratee 4

Step 1: No-Go - Apply petro to 1/2 of clamp assembly.

Step 2: No-Go - Install valve assembly backwards.

Step 3: No-Go - Insert bolts upside down--bolt head in wrong direction.

Step 4: No-Go - Gasket not seated or aligned properly.

Step 5: No-Go - Wrong measure; then, jiggles around but does not recheck measurement.

Step 6: No-Go - Fail to curl twisted safety-wire tail.

INSTALL BLEED AIR SYSTEM COLLECTOR BOWL

Ratee 5

Step 1: No-Go - Apply petro to 1/2 clamp assembly.

Step 2: Go

Step 3: Go

Step 4: No-Go - Position clamp improperly; i.e., not at 3:00 or 9:00 position.

Step 5: No-Go - Neglect to measure -- doesn't use straight edge at all.

Step 6: No-Go - Safety wire too sloppy; fail to align gasket.

INSTALL BLEED AIR SYSTEM COLLECTOR BOWL

Ratee 6

Step 1: Go

Step 2: Go

Step 3: Go

Step 4: Go

Step 5: Go

Step 6: Go

APPENDIX D:

PREDICTOR AND PROCESS VARIABLE MEASURES

BACKGROUND INFORMATION

Name: ____________________ Age: __________

SSN: ____________________ Sex: _________ (M/F)

Race: Black ___ White ___ Hispanic ___ Asian ___ American Indian ___

Other ___

Rank: E-__________

Years/Months in Air Force: ______

Years/Months as Mechanic
(including civilian and military experience): ______

Years/Months as Air Force Jet Engine Mechanic: ______

Years/Months experience with J79 Engine: ______

In the last six months, how many times have you installed a starter on a J79 Jet Engine? ______

In the last six months, how many times have you installed a collector bowl on a J79 Jet Engine? ______

Over the last 5 years, how often have you observed and formally evaluated the work of another mechanic?

Never ___ Once a Year ___ Once a Month ___ Once a Week ___ Daily ___

Over the last 5 years, how often have you observed and formally evaluated the work of a <u>non-mechanic</u>?

Never ___ Once a Year ___ Once a Month ___ Once a Week ___ Daily ___

How long has it been since you last observed and formally rated the performance of another mechanic?

Never Made Such Observations/Ratings ___ More Than One Year ___

One Year ___ One Month ___ One Week ___ Less Than One Week ___

Please Turn the Page and Continue

Below are a variety of questions about your background experiences. No two questions are exactly alike, so consider each statement carefully before answering. Please indicate how much you agree or disagree with these statements. Use the following 7-point scale:

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree Nor Disagree	Somewhat Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

As you can see, the higher your rating, the more you agree with the statement. Please answer as honestly as possible and make your rating (1-7) in the space provided at the left of each item.

_____ 1. It bothers me when other people are so concerned with little details of a job.

_____ 2. When working together with other mechanics, I usually ignore what they are doing and concentrate on my part of the job.

_____ 3. The kind of job I enjoy most is one that involves a lot of little steps.

_____ 4. Often I am able to complete a repair job that others cannot figure out how to do.

_____ 5. You should always make sure that the people around you are doing their part of the job correctly.

_____ 6. I do not like to point out errors in the work other people do when these errors are not really important.

_____ 7. I slept enough last night to feel good today.

_____ 8. Other mechanics I have worked with are more skilled than I am.

_____ 9. People think I am excessively concerned with unimportant details.

_____ 10. Whenever I work with other mechanics, no matter what their rank, I pay close attention to what they are doing to make sure that they do the job right.

_____ 11. I have a knack for taking things apart and putting them back together.

_____ 12. I usually like to get the major parts of a job done and let someone else handle the details.

_____ 13. Even when I have a chance to observe and correct other mechanics at work, I rarely do so.

_____ 14. I tend to be overcritical of other people's work.

_____ 15. Sometimes I feel as if I just don't have the skill or interest to be a first-rate mechanic.

For each of the following questions, place an "X" in the blank next to your preferred answer. Please mark <u>only one "X"</u> in response to each question.

16. What is the highest level of education you have completed?

_____ High school

_____ Trade/vocational school

_____ Two years of college

_____ Four years of college

17. What was your approximate high school grade average?

_____ A (3.5-4.0)

_____ high B (3.0-3.49)

_____ low B (2.5-2.99)

_____ high C (2.0-2.49)

_____ lower (less than 2.0)

18. Do you feel you are a good detail person?

_____ Definitely yes; I am very detail-oriented and attend closely to "nitty-gritties" of a task or job.

_____ I am probably about average on detail orientation.

_____ Not really; I tend to overlook small details or fail to do a really thorough job of attending to the details required on many tasks or jobs.

_____ Definitely no; I often miss important details on a task or job and I'm much better at things requiring little or no detail work.

19. How important do you feel it is to make an all-out effort on a task or job?

_____ not important at all

_____ not very important

_____ important

_____ very important

_____ extremely important

20. How well do you like to be around other people?

_____ I enjoy being with others very much; only rarely do I like to be by myself.

_____ I usually enjoy being around others, occasionally preferring to be by myself.

_____ I like being around other people sometimes and at other times I like to be by myself.

_____ I prefer being by myself and only occasionally enjoy being around other people.

21. How do you feel about a task or job requiring considerable attention to detail?

_____ enjoy it

_____ don't mind it

_____ dislike it

_____ thoroughly dislike it

22. How often do you find that your first impression of a person is the right one?

_____ always

_____ often

_____ occasionally

_____ rarely

_____ never

23. It doesn't bother me to put aside what I have been doing without finishing it.

_____ true

_____ false

24. I notice little things about a person or a situation that others overlook.

_____ this happens to me almost all the time

_____ this often happens to me

_____ this has happened to me several times, but I wouldn't say this is generally true of me

_____ this very seldom happens to me

_____ this never happens to me

25. Some people easily become involved in a task while there's seldom really "dig into" a task or job. How involved do you usually become in a task or job?

_____ I often have trouble sticking with it; other things almost always seem to come up to distract my attention.

_____ I sometimes become involved in a task or job that interests me greatly, but most of the time I quickly lose interest.

_____ I often become heavily involved in a task or job provided it's of interest to me.

_____ I almost always become engrossed in tasks or jobs.

For reasons of test security, the Orientation and Maze Tests cannot be included in this paper. These tests are controlled by the U.S. Army Research Institute, 5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600.

Adjective Check List

On the following pages is a list of 255 adjectives. Please read the adjectives quickly and place a check next to each adjective that you feel describes you. Do not worry about contradictions. Work quickly and do not spend too much time on any one adjective. Try to be frank, and place a check next to the adjectives that describe you as you really are, not as you would like to be.

___ 1. absent-minded
___ 2. active
___ 3. adaptable
___ 4. adventurous
___ 5. affectionate
___ 6. aggressive
___ 7. alert
___ 8. aloof
___ 9. ambitious
___ 10. anxious
___ 11. apathetic
___ 12. appreciative
___ 13. argumentative
___ 14. arrogant
___ 15. artistic
___ 16. assertive
___ 17. autocratic
___ 18. awkward
___ 19. bitter
___ 20. blustery
___ 21. boastful
___ 22. bossy
___ 23. calm
___ 24. capable
___ 25. careless
___ 26. cautious
___ 27. changeable
___ 28. cheerful
___ 29. civilized
___ 30. clear-thinking
___ 31. clever
___ 32. coarse
___ 33. cold
___ 34. commonplace
___ 35. complaining
___ 36. complicated
___ 37. conceited
___ 38. confident
___ 39. confused
___ 40. conscientious
___ 41. conservative
___ 42. considerate
___ 43. contented
___ 44. conventional
___ 45. cool
___ 46. cooperative
___ 47. courageous
___ 48. curious
___ 49. cynical
___ 50. daring
___ 51. defensive
___ 52. deliberate
___ 53. demanding
___ 54. dependable
___ 55. dependent
___ 56. determined
___ 57. dignified
___ 58. discreet
___ 59. disorderly
___ 60. dissatisfied
___ 61. distractible
___ 62. distrustful
___ 63. dominant
___ 64. dull
___ 65. easy-going
___ 66. efficient
___ 67. egotistical
___ 68. emotional
___ 69. energetic
___ 70. enterprising
___ 71. enthusiastic
___ 72. evasive
___ 73. excitable
___ 74. fair-minded
___ 75. fault-finding
___ 76. fearful
___ 77. fickle
___ 78. forceful
___ 79. foresighted
___ 80. forgetful
___ 81. forgiving
___ 82. formal
___ 83. frank
___ 84. friendly
___ 85. fussy
___ 86. generous
___ 87. gentle
___ 88. gloomy
___ 89. good-natured
___ 90. hard-headed
___ 91. hard-hearted
___ 92. hasty
___ 93. headstrong
___ 94. healthy
___ 95. helpful
___ 96. high-strung
___ 97. honest
___ 98. hostile
___ 99. humorous
___ 100. hurried
___ 101. idealistic
___ 102. imaginative
___ 103. impatient
___ 104. impulsive
___ 105. independent
___ 106. indifferent
___ 107. individualistic
___ 108. industrious
___ 109. informal
___ 110. ingenious
___ 111. inhibited
___ 112. insightful
___ 113. intelligent
___ 114. interests narrow
___ 115. interests wide
___ 116. intolerant
___ 117. inventive
___ 118. irritable
___ 119. jolly
___ 120. kind
___ 121. lazy
___ 122. leisurely
___ 123. logical
___ 124. loud
___ 125. loyal
___ 126. mannerly

Go on to next page

___127. mature
___128. meek
___129. methodical
___130. mild
___131. mischievous
___132. moderate
___133. modest
___134. moody
___135. nagging
___136. natural
___137. nervous
___138. noisy
___139. obliging
___140. opinionated
___141. opportunistic
___142. optimistic
___143. organized
___144. original
___145. outgoing
___146. outspoken
___147. painstaking
___148. patient
___149. peaceable
___150. persevering
___151. persistent
___152. pessimistic
___153. planful
___154. pleasant
___155. pleasure-seeking
___156. poised
___157. polished
___158. practical
___159. praising
___160. precise
___161. preoccupied
___162. progressive
___163. quarrelsome
___164. quick
___165. quiet
___166. rational
___167. realistic
___168. reasonable
___169. rebellious
___170. reckless
___171. reflective
___172. relaxed
___173. reliable
___174. resentful
___175. reserved
___176. resourceful
___177. responsible
___178. restless
___179. retiring
___180. rigid
___181. robust
___182. sarcastic
___183. self-centered
___184. self-confident
___185. self-controlled
___186. self-denying
___187. self-pitying
___188. self-punishing
___189. self-seeking
___190. selfish
___191. sensitive
___192. sentimental
___193. serious
___194. severe
___195. sharp-witted
___196. show-off
___197. shrewd
___198. shy
___199. silent
___200. simple
___201. sincere
___202. slow
___203. sly
___204. smug
___205. sociable
___206. soft-hearted
___207. sophisticated
___208. spendthrift
___209. spontaneous
___210. spunky
___211. stable
___212. steady
___213. stern
___214. stingy
___215. strong
___216. stubborn
___217. suggestible
___218. superstitious
___219. suspicious
___220. sympathetic
___221. tactful
___222. tactless
___223. talkative
___224. temperamental
___225. tense
___226. thorough
___227. thoughtful
___228. thrifty
___229. timid
___230. tolerant
___231. tough
___232. trusting
___233. unaffected
___234. unambitious
___235. unassuming
___236. unconventional
___237. undependable
___238. understanding
___239. unemotional
___240. unexcitable
___241. unfriendly
___242. uninhibited
___243. unkind
___244. unrealistic
___245. unscrupulous
___246. unselfish
___247. vindictive
___248. versatile
___248. warm
___250. wary
___251. weak
___252. wise
___253. withdrawn
___254. witty
___255. worrying

Post-Rating Questionnaire

Below are a variety of questions about your experience rating the videotape. No two questions are exactly alike, so consider each statement carefully before answering. Please indicate how much you agree or disagree with these statements. Use the following 7-point scale:

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree Nor Disagree	Somewhat Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

As you can see, the higher your rating, the more you agree with the statement. Please answer as honestly as possible and make your rating (1-7) in the space provided at the left of each item.

____ 1. I probably did better rating the first few performances than I did rating the last few performances.

____ 2. Had the lighting been better or the position of the camera been different, I could have seen more clearly whether or not the mechanic performed the tasks correctly.

____ 3. I really wanted to do well on this task.

____ 4. I felt that I should give the mechanic on the videotape a no-go whenever he made a mistake, even if the mistake was fairly minor.

____ 5. There are several good ways to do the two tasks, and the Tech Orders only presents one of these ways.

____ 6. Sometimes I couldn't tell how the mechanic was supposed to do the job.

____ 7. If I knew an equally good or better way to do the tasks on the videotape, I ignored the Tech Orders and relied on my own judgment.

____ 8. I was able to see small details in the videotaped performances that others may have been unable to see.

____ 9. I was very relieved when the videotape was finally over.

____ 10. Because of my location in the room, I was unable to see the television monitor as well as the others could.

____ 11. To tell you the truth, I really didn't care how accurate my ratings were.

____ 12. Even after reading the Tech Orders carefully, it was sometimes unclear how the person should do the job.

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree Nor Disagree	Somewhat Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

____ 13. I really wanted to find as many mistakes as possible in the videotaped performances.

____ 14. I had no trouble seeing what the mechanics were doing on the videotape.

____ 15. When rating the performances on the videotape, I was always sure how the person should do the job.

____ 16. When rating the performances on the tape, I was more concerned with <u>whether</u> the mechanic got the job done than with <u>how</u> he got the job done.

____ 17. If I was not sure how the mechanic should do the task, I looked first to the Tech Orders I was given.

____ 18. Sometimes the other people in the room made noises or said something that gave me a clue that a mistake had been made in the videotape.

____ 19. If the mechanic on the tape did something wrong, but I thought it was minor, I gave him a "go" anyway.

____ 20. Although I know I don't have to, I would gladly rate more videotaped performances right now.

____ 21. Even when the Tech Orders were unclear or incomplete, I was able to determine the right way to do these tasks.

____ 22. During each videotaped performance, I looked for every little detail that would help me determine whether or not the performance was done correctly.

____ 23. Most people would say that there are 36 hours in a day.

____ 24. I tried to follow the Tech Orders as carefully as possible to determine whether or not the mechanic performed each step correctly.

____ 25. I often found myself thinking about other things I have to do rather than thinking about the rating task.

____ 26. I might have been more accurate if I could have seen more clearly what the mechanics were doing on the videotape.

____ 27. Even after the first few performances, when others may have been tired or bored, I continued to rate the performances as accurately as I could.

APPENDIX E:

CODING SCHEME FOR REASONS FOR GO/NO-GO RATINGS

Condensed List of Reasons (Aggregated over Ratees and Videotapes)

Starter Installation

Step 1

0 No Reason/Nothing Wrong
1 General: Did it wrong
2 Too much/a lot of grease/didn't wipe off excess
3 Too little grease
4 Should use acid brush/improper application method
5 Had a watch on
6 Didn't put grease on starter shaft/
Put grease in wrong place
7 Didn't apply grease to all splines/
Didn't check to see if all
splines were greased
8 Didn't cover all splines evenly
9 Should wrap excess wire around pliers
10 Wearing a watch (see 5; mistake)
11 Used his hands
12 Spline grease dirty
13 Wrong type of grease
14 Wearing jewelry

Step 2

0 No Reason
1 General: Clamp hung wrong/clamp not completely installed
2 Put clamp in wrong place
(Should have put on engine/
should have put on starter/
should place clamp on adapter)
3 Didn't latch coupling/lock clamp/latch properly/secure it
4 Damaged clamp/wrong adapter/t-bolt bent
5 Clamp in wrong position (e.g., 9 o'clock not 3 o'clock)
6 Clamp hung too tightly/shouldn't be bolted
7 Wrong side of adapter plate
8 Too slow
9 T-bolt not installed properly
10 Easier in other position
11 Didn't cement clamp

Step 3

0 No Reason
1 General: Did it wrong
2 Didn't make sure starter was
on flange properly/
Didn't work starter onto flange/
Starter not engaged/
Didn't rotate
Starter not seated
3 Let go of starter too soon/
before torquing, before step 4/
Didn't support starter/
Starter supported by shaft
4 Starter in wrong position/
Starter not positioned/
Starter not indexed/
Breach in wrong position
5 Didn't thread coupling nut/nut not tight enough
6 Had to unlock coupling latch
7 Clumsy
8 Coupling not positioned properly
9 Improper use of tools/used speed handle wrong
10 Starter not under locking edge of coupling band

Step 4

0 No Reason
1 General: Did it wrong
2 Clamp in wrong position/not indexed properly
3 Clamp not seated/not seated enough/didn't use mallet
Didn't check for proper seating
4 Tightened coupling nut before seating it/
job done in wrong order
5 T-bolt positioned improperly in clamp/
T-bolt installed improperly
6 Coupling damaged/bad clamp
7 Used washer on clamp
8 Didn't torque nut/tighten it/
Clamp not tight
9 Didn't keep tensions on splines until torqued
10 Didn't support starter
11 Overtorqued coupling
12 Shouldn't use speed wrench
13 Didn't check for proper position

Step 5

0 No Reason
1 General: Did it wrong
2 Handled torque wrench wrong
3 Slow
4 Jerky/Clumsy
5 Overtorqued
6 Didn't seat V-band/didn't tap
7 Didn't keep tension on splines until torqued
8 Not wet-torqued
9 Didn't check torque wrench prior to use
10 Undertorqued/torque did not click
11 Double torqued

Step 6

0 No Reason
1 General: Did it wrong
2 Overtorqued nut
3 Handled torque wrench wrong/improper torque
4 Safety nut was old/wrong kind
5 Wrong tool was used
6 Didn't torque nut/backup nut/safety nut/
Didn't use torque wrench
7 Slow
8 Not wet-torqued
9 Seated nut without torquing it
10 Undertorqued
11 Didn't hold first nut when turning second
12 Double torqued
13 Didn't check torque

Step 7

0 No Reason
1 General: Did it wrong
2 General: Improper safety procedures/
bad safety/(1)wrong type of safety wire
3 Pigtail backwards/twisted wrong way/backwards/reversed
4 Safety wire too long
5 Loose safety
6 Safety wire neutral
7 Plug not finger tight
8 Used wrong tool/should use dikes to cut wire
should use needle nose for pigtail
9 Didn't go through hole/not anchored/not safetied to anything
10 Too many twists/too tight
11 Safety undertwisted
12 Used wrong type of safety wire
13 Didn't seat plug/didn't do it well enough
14 Undid safety
15 Plug only finger tight
15 Overtorqued cannon plug
16 Mating holes and pins not checked

Collector Bowl Installation

Step 1

0 No Reason
1 General: Did it wrong
2 Too much lubricant
3 Too little lubricant
4 Should use brush
5 Applied to only one side/not all of clamp
6 Wrong side of clamp
7 Didn't spread it on

Step 2

0 No Reason
1 General: Did it wrong
2 Left out gasket
3 Valve positioned wrong/
Index down(?)
4 Didn't index valve assembly/
Valve not checked for alignment/
Didn't check valve index
5 Did not put both gaskets on
6 Rubbed fingers on gasket surface
7 Broken flapper
8 Bad technique
9 Gasket not seated properly

Step 3

0 No Reason
1 General: Did it wrong
2 Bolts upside down/positioned improperly/backwards
3 Back coupling/stub duct up too soon
4 No second gasket/no gasket on back side
5 Clumsy
6 Missing washer/no washer on back side of coupling
7 Could have cut or pinched (something)
8 Valve positioned wrong/rotated during installation
9 Poor technique
10 Gasket out of groove/not aligned
11 Duct not seated correctly to bowl at first
12 Clamp not seated
13 Front and rear clamps not properly aligned

Step 4

0 No Reason
1 General: Did it wrong/
Not following Technical Order sequence
2 Gasket in wrong place when coupling installed/
Didn't ensure proper seating of gasket/
Gasket not aligned/installed gasket incorrectly
3 No gasket/gasket missing/seal missing
4 Installed clamp first (before gasket)
5 Installed nut before positioning coupling/
Tightened nut
6 Coupling/clamp positioned improperly
7 Had no running torque/didn't torque
8 Wrong parts: Clamp not serviceable/
wrong clamp bolt/did not need washer
9 Didn't place T-bolt through eye
10 Used washer in clamp
11 Didn't rotate to upper portion of duct
12 Used a spacer/washer not needed (see 10)
13 Should have pre-safetied clamp
14 Didn't seat clamp
15 Bolt upside down
16 Didn't lube clamp
17 Didn't thread nut
18 Bad technique
19 Spacer installed improperly
20 Tightened nut instead of just threading it/
shouldn't tighten nut (see 5)

Step 5

0 No Reason
1 General: Did it wrong/not following procedures
2 Measurement off/duct flange not aligned
3 Didn't measure/recheck measurement
4 Ruler angled/
didn't position at top of duct flange
5 Wrong torque nut
6 Should use rubber mallet to position duct
7 Gasket is damaged
8 2, 3, or 4 (can't tell)
9 Measurement taken at wrong place
10 Measured, then moved it

Step 6

0 No Reason
1 General: Did it wrong/wrong sequence
2 General: Safety wire wrong/not very good
Improper safety wire
3 Didn't torque first nut/first backup nut/nut not torqued right
4 Overtorqued nut/backup nut
first nut/clamp overtorqued
5 Cross-threaded first nut
6 Didn't install backup nut/no backup nut
7 Safety wire over-twisted/
too many twists/too tight
8 Safety wire not neat
9 Undid a safety twist
10 Nuts double torqued/torqued nuts together
11 Safety-wired before installing backup nut/
Installed safety nut in wrong order/
12 Didn't pigtail safety wire/
Safety wire not bent/curled/
13 Not enough threads showing on coupling
14 Should have used extension while torquing
15 Didn't torque backup nut
16 Loose safety/not completely twisted
17 Didn't seat duct/clamp
18 Duct positioned wrong/didn't ensure correct position
19 Clamp positioned wrong
20 Pigtail backwards
21 Nut not wet-torqued
22 Should wrap excess wire around pliers
23 Safety nut not installed (see 6)
24 Improper nut installation
25 Improper torque
26 Wrong safety wire
27 Didn't hold first nut while torquing the second
28 Shouldn't leave wire loose while installing the backup nut
29 Poor/wrong handling of pliers/speed handle
30 Poor technique
31 Slow
32 Didn't double loop safety wire at end of twist

APPENDIX F:

DETAILS OF EXPERT RATER RESULTS

Expert Rater Results

Experts	Starter 1	2	3	4	5	6	7	Collector Bowl 1	2	3	4	5	6
Ratee 1													
1	G	G	G	G	G	G	G	G	N	G	G	G	N
2	G	G	G	G	G	G	G	G	N	G	G	G	N
3	G	G	N	G	G	G	G	G	N	G	G	G	N
4	G	G	N	G	G	G	G	G	N	G	G	G	N
5	G	G	G	G	G	N	G	G	N	G	G	G	N
Consensus	G	G	N	G	G	G	G	G	N	G	G	G	N
Intended	G	G	G	G	G	G	G	G	G	G	G	G	N
Ratee 2													
1	G	G	G	N	G	G	G	G	N	G	N	N	N
2	G	G	N	N	G	G	N	G	N	G	N	N	N
3	G	G	G	N	G	N	G	G	N	G	N	N	N
4	G	G	G	N	G	G	N	G	N	G	N	N	G
5	G	G	N	N	G	N	G	G	N	G	N	N	N
Consensus	G	G	N	N	G	G	N	G	N	G	N	N	N
Intended	G	G	G	N	G	G	N	G	N	G	N	N	N
Ratee 3													
1	G	N	N	N	G	G	N	G	G	G	G	N	G
2	G	N	N	N	N	G	G	G	G	G	G	G	N
3	G	N	N	G	N	G	G	G	G	G	G	G	N
4	G	N	N	N	N	G	G	G	G	G	G	G	N
5	G	N	N	G	N	G	G	N	G	G	G	G	N
Consensus	G	N	N	N	N	G	G	C	G	G	G	G	N
Intended	G	N	N	N	N	G	G	G	N	G	G	G	N

Note: G = go rating; N = no-go rating.

Experts	Starter 1	2	3	4	5	6	7	Collector Bowl 1	2	3	4	5	6
Ratee 4													
1	N	G	G	G	N	G	N	N	N	N	G	N	N
2	N	G	G	G	N	G	N	N	N	N	N	N	N
3	G	G	G	G	N	G	N	N	N	N	N	N	G
4	G	G	G	G	N	G	N	N	N	N	G	N	N
5	N	G	G	G	N	G	N	N	N	N	N	N	N
Consensus	N	G	G	G	N	G	N	N	N	N	N	N	N
Intended	N	G	G	G	N	G	N	N	N	N	G	N	N
Ratee 5													
1	N	G	G	N	N	N	G	N	G	G	G	N	N
2	N	G	G	N	N	N	G	N	N	G	N	N	N
3	N	G	G	N	G	N	G	N	N	G	N	N	N
4	N	G	G	N	G	N	G	N	N	G	N	N	N
5	N	G	G	N	N	N	G	N	G	G	N	N	N
Consensus	N	G	G	N	N	N	G	N	G	G	N	N	N
Intended	N	G	G	N	N	N	G	N	G	G	N	N	N
Ratee 6													
1	G	G	G	G	G	G	G	N	G	G	G	G	N
2	G	G	G	G	G	G	G	G	G	G	N	N	G
3	G	G	N	N	G	G	G	G	G	G	G	G	G
4	G	G	G	G	G	G	G	G	G	G	G	G	G
5	G	G	G	N	G	G	G	G	G	G	G	G	N
Consensus	G	G	G	N	G	G	G	G	G	G	G	G	G
Intended	G	G	G	N	G	G	G	G	G	G	G	G	G

Note: G = go rating; N = no-go rating.